DOI:
10.1109/TSIPN.2021.3070712

Document Version
Peer reviewed version

Collaborative Cloud and Edge Mobile Computing in C-RAN Systems with Minimal End-to-End Latency

Seok-Hwan Park, *Member*, *IEEE*, Seongah Jeong, *Member*, *IEEE*, Jinyeop Na, *Student Member*, *IEEE*, Osvaldo Simeone, *Fellow*, *IEEE*, and Shlomo Shamai (Shitz), *Life Fellow*, *IEEE*

***Abstract*—Mobile cloud and edge computing protocols make it possible to offer computationally heavy applications to mobile devices via computational offloading from devices to nearby edge servers or more powerful, but remote, cloud servers. Previous work assumed that computational tasks can be fractionally offloaded at both cloud processor (CP) and at a local edge node (EN) within a conventional Distributed Radio Access Network (D-RAN) that relies on non-cooperative ENs equipped with one-way uplink fronthaul connection to the cloud. In this paper, we propose to integrate collaborative fractional computing across CP and ENs within a Cloud RAN (C-RAN) architecture with finite-capacity two-way fronthaul links. Accordingly, tasks offloaded by a mobile device can be partially carried out at an EN and the CP, with multiple ENs communicating with a common CP to exchange data and computational outcomes while allowing for centralized precoding and decoding. Unlike prior work, we investigate joint optimization of computing and communication resources, including wireless and fronthaul segments, to minimize the end-to-end latency by accounting for a two-way uplink and downlink transmission. The problem is tackled by using fractional programming (FP) and matrix FP. Extensive numerical results validate the performance gain of the proposed architecture as compared to the previously studied D-RAN solution.**

***Index Terms*—Mobile cloud computing, edge computing, C-RAN, constrained fronthaul, end-to-end latency minimization, (matrix) fractional programming.**

I. Introduction

Mobile cloud and edge computing techniques enable computationally heavy applications such as gaming and augmented reality (AR) by offloading computation tasks from battery-limited mobile user equipments (UEs) to cloud or edge servers which are located respectively at cloud processor (CP) or edge nodes (ENs) of a cellular architecture [1]–[7]. In systems with both cloud and edge computing capabilities, computation tasks can be opportunistically offloaded either to ENs or to the CP [8]. For example, it may be desirable to offload latency-insensitive and computationally heavy tasks to a CP, while relatively light tasks with more stringent latency constraints can be offloaded to edge servers in ENs.

The optimization of the offloading decision policy was studied in [9], [10] by focusing on the application layer and without including constraints imposed by the Radio Access Network (RAN). To the best of our knowledge, reference [3] for the first time studied the *joint* optimization of computation and communication resources for mobile wireless edge computing systems, with follow-up works including [4]. Both papers [3], [4] aimed at minimizing energy expenditure under constraints on the end-to-end latency that encompass the contributions of both communication and computation. While [3] accounts only for uplink transmission, reference [4] also includes the contribution of downlink communication, which is required to feed back the results of the remote computations. To overcome the inherent non-convexity of the resulting optimization problems, the authors in [3], [4] applied successive convex approximation (SCA) [11], [12], which efficiently finds a locally optimal solution for constrained non-convex problems. Extensions in [13], [14] studied edge computing-based AR applications [13] and edge computing via an unmanned aerial vehicle (UAV) mounted cloudlet [14].

In a system with both *cloud and edge computing* capabilities, computation tasks can be partially offloaded to CP and ENs [8]. Reference [8] tackled the problem of jointly optimizing communication and computational resources with the goal of minimizing a weighted sum of per-UE end-to-end latency metrics within a distributed RAN (D-RAN) architecture [15, Sec. III]. The authors in [8] developed closed-form solutions for optimal resource allocation and task splitting ratios by focusing on the design of uplink communication from UEs to ENs and CP while assuming orthogonal time-division multiple access (TDMA) on wireless access uplink channel and a fixed allocation of fronthaul capacity across the UEs. Reference [16] also addressed the design of the task splitting ratios under the assumption that the task of each UE can be split into multiple subtasks that are offloaded to multiple ENs.

In a D-RAN, ENs perform local signal processing for channel encoding and decoding. Thus, the overall performance

S.-H. Park was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grants funded by the Ministry of Education [NRF-2019R1A6A1A09031717, 2021R1C1C1006557]. The work of S. Jeong was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-0-01787) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation). This work was also supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement Nos. 694630 and 725731).

S.-H. Park is with the Division of Electronic Engineering and the Future Semiconductor Convergence Technology Research Center, Jeonbuk National University, Jeonju 54896, Korea (email: seokhwan@jbnu.ac.kr).

S. Jeong is with the School of Electronics Engineering, Kyungpook National University, Daegu 14566, Korea (email: seongah@knu.ac.kr).

J. Na is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea (email: wlsduq37@kaist.ac.kr).

O. Simeone is with King's Communication, Learning and Information Processing (kclip) Lab, the Centre for Telecommunications Research, Department of Engineering, King's College London, London WC2R 2LS, U.K (email: osvaldo.simeone@kcl.ac.uk).

S. Shamai is with the Department of Electrical and Computer Engineering, Technion, Haifa 3200003, Israel (email: sshlomo@ee.technion.ac.il).

can be degraded by interference in dense networks. In this paper, we propose integrating collaborative fractional cloud-edge offloading within a cloud radio access network (C-RAN) architecture [17], while accounting for the contributions of both uplink and downlink. In a C-RAN, as illustrated in Fig. 1, joint signal processing, in the form of cooperative precoding and detection, at the CP enables effective interference management. Unlike the case of D-RANs, the design of C-RAN systems entails the additional challenge of optimizing the use of ENs-CP *fronthaul* links [18]–[20]. In this regard, we note that, although fronthaul constraints were also considered in [8] for the design within a D-RAN system, a simple data forwarding model was assumed with fixed capacity allocation among the UEs. In [21], the authors tackled the optimization of functional split for collaborative computing systems equipped with a packet-based fronthaul network. However, it was assumed in [21] that the physical-layer (PHY) functionalities, which include channel encoding and decoding, are located only at ENs. In [22], the authors addressed the task allocation and traffic path planning problem for a C-RAN system under the assumption that the service latency consists of task processing delay and path delay only on fronthaul links.

In this work, we address the optimization of C-RAN signal processing for the purpose of enabling collaborative cloud and edge mobile computing with minimal end-to-end two-way latency. We proceed by first reviewing the design of collaborative cloud and edge computing system within a D-RAN architecture. Unlike [8], [23], which considered one-way uplink design with inter-UE TDMA and fixed fronthaul capacity allocation, we address the design of two-way communications with both TDMA and non-orthogonal multiple access strategies and we treat the fronthaul capacity allocation as optimization variables. Then, we address the design of C-RAN system for collaborative offloading. For all the design problems, we consider the criterion of minimizing two-way end-to-end latency for computation offloading as in [8], [24]–[26]. To tackle the formulated problems, which turn out to be non-convex, we adopt fractional programming (FP) and matrix FP [27], [28]. We present extensive numerical results that confirm the convergence of the proposed optimization algorithms, the advantages of C-RAN architecture as compared to D-RAN [8], and the impact of collaborative cloud and edge computing on latency with C-RAN.

The paper is organized as follows. In Sec. II, we describe the system model including the computational tasks, computational capabilities, wireless channel and fronthaul transmission models. In Sec. III, we discuss the design of collaborative cloud and edge mobile computing system within the D-RAN architecture, and the design for a C-RAN system is discussed in Sec. IV. We provide extensive numerical results in Sec. V to validate the performance gain of the proposed architecture as compared to the D-RAN solution. We conclude the paper in Sec. VI.

Notations: We denote the set of all $M \times N$ complex matrices by $\mathbb{C}^{M\times N}$. The notation $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ indicates that $\mathbf{x}$ is a column vector following circularly symmetric complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Omega}$. We also use the notation $I(\mathbf{x};\mathbf{y})$ to represent the mutual information between random vectors $\mathbf{x}$ and $\mathbf{y}$. A block diagonal matrix, whose diagonal blocks are given as $\mathbf{A}_1, \ldots, \mathbf{A}_L$, is denoted by $\text{diag}(\{\mathbf{A}_l\}_{l\in\{1,\ldots,L\}})$. Lastly, $\mathbb{E}[\cdot]$ represents the expectation operator, and $||\mathbf{x}||$ denotes the Euclidean 2-norm of a vector $\mathbf{x}$.

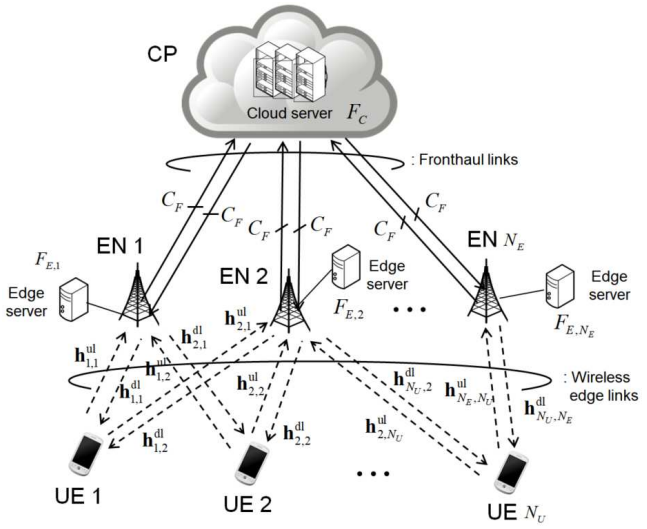


Figure 1. Illustration of collaborative cloud and edge mobile computing system within C-RAN architecture.

II. System Model

As illustrated in Fig. 1, we consider a collaborative cloud and edge mobile computing system, in which N_U single-antenna mobile UEs offload their computational tasks to a network consisting of N_E ENs and a CP. In order to exchange computational input information, the UEs communicate with the ENs over a wireless uplink channel, and each EN is connected to the CP through dedicated fronthaul link of finite capacity C_F^{ul} bits per second (bps). For communication in the reverse direction from CP to each EN, the fronthaul has capacity of C_F^{dl} bps, and the ENs transmit to the UEs in a wireless downlink channel. For convenience, we define the sets $\mathcal{N}_U \triangleq \{1, 2, \ldots, N_U\}$ and $\mathcal{N}_E \triangleq \{1, 2, \ldots, N_E\}$ of indices of UEs and ENs, respectively. We denote the number of antennas of EN i as $n_{E,i}$, and the number of all ENs' antennas is $n_E = \sum_{i\in\mathcal{N}_E} n_{E,i}$. The bandwidths of uplink and downlink channels are W^{ul} and W^{dl}, respectively, which are measured in Hz.

A. Computational Tasks and Collaborative Computing Model

As in [4], [8], we assume that the UEs have limited computing powers, and hence offload their whole tasks to ENs or CP without local processing. We define $b_{I,k}$ and $b_{O,k}$ as the numbers of input and output bits for the task of UE k. We assume that V_k CPU cycles are required to process one bit of the task of UE k so that the task of UE k requires $b_{I,k}V_k$ CPU cycles in total. The computing powers of each EN i and CP are denoted by $F_{E,i}$ and F_C, respectively, whose units are CPU cycles per second.

For each UE k, we allow for a collaborative cloud and edge computing [4], [8]. This means that a part of the task of UE k is processed by a predetermined EN i_k, while the rest of the task is offloaded to the CP. We define a variable $c_k \in [0,1]$ which controls the fraction of the task of UE k that is processed by EN i_k. Accordingly, EN i_k receives the input information of $c_k b_{I,k}$ bits from UE k, runs $c_k b_{I,k} V_k$ CPU cycles, and reports the resulting output information of $c_k b_{O,k}$ bits back to UE k. Similarly, the CP receives $(1-c_k)b_{I,k}$ input bits from UE k, runs $(1-c_k)b_{I,k}V_k$ CPU cycles, and sends $(1-c_k)b_{O,k}$ output bits to UE k.

We define $\mathcal{N}_{U,i}$ as the set of UEs that are associated with EN i, i.e.,

$$\mathcal{N}_{U,i} = \left\{k \in \mathcal{N}_U | i_k = i\right\}. \tag{1}$$

Therefore, if we denote as $F_{E,i,k}$ the computing power of EN i assigned for UE k, the variables $F_{E,i,k}$, $k \in \mathcal{N}_{U,i}$, are subject to the constraint

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} F_{E,i,k} \leq F_{E,i}. \tag{2}$$

The edge computation latency $\tau^{\text{exe}}_{E,i,k}$ for UE k at EN i with $k \in \mathcal{N}_{U,i}$ is given as

$$\tau^{\text{exe}}_{E,i,k} = \frac{c_k b_{I,k} V_k}{F_{E,i,k}}. \tag{3}$$

Similarly, denoting the computing power allocated to UE k by the CP as $F_{C,k}$, the variables $F_{C,k}$, $k \in \mathcal{N}_U$, should satisfy the constraint

$$\sum\nolimits_{k\in\mathcal{N}_U} F_{C,k} \leq F_C. \tag{4}$$

The cloud computing latency $\tau^{\text{exe}}_{C,k}$ for UE k at the CP is given as

$$\tau^{\text{exe}}_{C,k} = \frac{(1-c_k) b_{I,k} V_k}{F_{C,k}}. \tag{5}$$

B. Wireless Channel Model for Edge Link

Assuming the flat fading channel model for both the uplink and downlink wireless edge links, the received signal vector $\mathbf{y}^{\text{ul}}_i \in \mathbb{C}^{n_{E,i}\times 1}$ of EN i on the uplink is given as

$$\mathbf{y}^{\text{ul}}_i = \sum\nolimits_{k\in\mathcal{N}_U} \mathbf{h}^{\text{ul}}_{i,k} x^{\text{ul}}_k + \mathbf{z}^{\text{ul}}_i, \tag{6}$$

where $\mathbf{h}^{\text{ul}}_{i,k} \in \mathbb{C}^{n_{E,i}\times 1}$ denotes the channel vector from UE k to EN i; $x^{\text{ul}}_k \in \mathbb{C}^{1\times 1}$ indicates the transmit signal of UE k; and $\mathbf{z}^{\text{ul}}_i \sim \mathcal{CN}(\mathbf{0}, \sigma^2_{z,\text{ul}}\mathbf{I})$ is the additive noise vector. Similarly, the received signal $y^{\text{dl}}_k \in \mathbb{C}^{1\times 1}$ of UE k on the downlink can be written as

$$y^{\text{dl}}_k = \sum\nolimits_{i\in\mathcal{N}_E} \mathbf{h}^{\text{dl}H}_{k,i} \mathbf{x}^{\text{dl}}_i + z^{\text{dl}}_k, \tag{7}$$

where $\mathbf{h}^{\text{dl}}_{k,i} \in \mathbb{C}^{n_{E,i}\times n_{E,i}}$ represents the channel vector from EN i to UE k; $\mathbf{x}^{\text{dl}}_i \in \mathbb{C}^{n_{E,i}\times 1}$ denotes the transmit signal vector of EN i; and $z^{\text{dl}}_k \sim \mathcal{CN}(0, \sigma^2_{z,\text{dl}})$ denotes the additive noise.

The transmit powers of each UE k and EN i are limited as

$$\mathbb{E}\left[|x^{\text{ul}}_k|^2\right] \leq P^{\text{ul}}, \text{ and} \tag{8}$$

$$\mathbb{E}\left[||\mathbf{x}^{\text{dl}}_i||^2\right] \leq P^{\text{dl}}, \tag{9}$$

Symbol	Meaning
N_U, N_E	Numbers of UEs and ENs
$\mathcal{N}_U$, $\mathcal{N}_E$	Sets of UEs and ENs' indices
$n_{E,i}$	Number of antennas of EN i
C^{ul}_F, C^{dl}_F	Capacity of uplink ad downlink fronthaul links
W^{ul}, W^{dl}	Bandwidths of uplink and downlink channels
$b_{I,k}$, $b_{O,k}$	Numbers of input and output bits for UE k
V_k	Number of CPU cycles per input bit for UE k
$F_{E,i}$, F_C	CPU frequencies of EN i and CP
c_k	Fraction of the task of UE k processed by EN i_k
$\mathcal{N}_{U,i}$	Set of UEs associated with EN i
P^{ul}, P^{dl}	Maximum transmit powers of each UE and EN
$\sigma^2_{z,\text{ul}}$, $\sigma^2_{z,\text{dl}}$	Noise powers per receive antenna at ENs and UEs
$\text{SNR}^{\text{ul}}_{\max}$, $\text{SNR}^{\text{dl}}_{\max}$	Maximum SNRs of uplink and downlink channels
$\mathbf{h}^{\text{ul}}_{i,k}$, $\mathbf{h}^{\text{dl}}_{k,i}$	Uplink & downlink channels btw. UE k and EN i
$\mathbf{y}^{\text{ul}}_i$, y^{dl}_k	Received signals of EN i and UE k
x^{ul}_k, $\mathbf{x}^{\text{dl}}_i$	Transmitted signals of UE k and EN i
$\mathbf{z}^{\text{ul}}_i$, z^{dl}_k	Noise signals at EN i and UE k

Table I: Table summarizing important symbols used throughout the paper

where P^{ul} and P^{dl} represent the maximum transmit powers at each UE and EN, respectively. We define the maximum signal-to-noise ratios (SNRs) of the uplink and downlink channels as $\text{SNR}^{\text{ul}}_{\max} = P^{\text{ul}}/\sigma^2_{z,\text{ul}}$ and $\text{SNR}^{\text{dl}}_{\max} = P^{\text{dl}}/\sigma^2_{z,\text{dl}}$, respectively. The symbols described in this section are summarized in Table I.

III. Optimization for the D-RAN Architecture

In this section, we discuss the design of the collaborative cloud and edge mobile computing system under a D-RAN architecture [15, Sec. III]. Unlike [8], which considered one-way uplink design with inter-UE TDMA and fixed fronthaul capacity allocation, we address the design of two-way communications with both TDMA and non-orthogonal multiple access strategies while treating the fronthaul capacity allocation as optimization variables.

In D-RAN, each EN i locally decodes the uplink input information transmitted by the associated UEs $\mathcal{N}_{U,i}$ without cooperating with nearby ENs. Also, in the downlink, the computation output information for UEs $\mathcal{N}_{U,i}$ is solely encoded and transmitted by the serving EN i. We discuss the designs with orthogonal TDMA and non-orthogonal multiple access strategies in Sec. III-A and III-B, respectively.

A. Orthogonal TDMA

With TDMA, N_U UEs communicate with N_E ENs on the wireless edge link while being assigned different time slots so that there is no inter-UE interference on wireless channel. We define $u^{\text{ul}}_k \in [0,1]$ and $u^{\text{dl}}_k \in [0,1]$ as the uplink and downlink time fractions allocated to UE k. Thus, the defined fraction variables $\mathbf{u} \triangleq \{u^{\text{ul}}_k, u^{\text{dl}}_k\}_{k\in\mathcal{N}_U}$ should satisfy the constraint

$$\sum\nolimits_{k\in\mathcal{N}_U} u^{\text{ul}}_k = \sum\nolimits_{k\in\mathcal{N}_U} u^{\text{dl}}_k = 1. \tag{10}$$

In the uplink, UE k transmits a baseband signal which encodes the input information for its task. Assuming that Gaussian channel codebooks are used, the transmitted signal x^{ul}_k of UE k is distributed as $x^{\text{ul}}_k \sim \mathcal{CN}(0, p^{\text{ul}}_k)$. Since there is no co-channel interference with orthogonal TDMA, the transmit power p_k of UE k is set to $p^{\text{ul}}_k = P^{\text{ul}}$ without loss of optimality.

With the described transmission model, the achievable data rate R_k^{ul} between UE k and EN i in the uplink channel is given as $R_k^{\text{ul}} = u_k^{\text{ul}} W^{\text{ul}} I(x_k^{\text{ul}}; \mathbf{y}_i^{\text{ul}})$, where the mutual information $I(x_k^{\text{ul}}; \mathbf{y}_i^{\text{ul}})$ is calculated as

$$I\left(x_k^{\text{ul}}; \mathbf{y}_i^{\text{ul}}\right) = \log_2\left(1 + \left(P^{\text{ul}}/\sigma_{z,\text{ul}}^2\right) \left\|\mathbf{h}_{i,k}^{\text{ul}}\right\|^2\right). \tag{11}$$

The uplink latency $\tau_{E,k}^{\text{ul}}$ on the wireless edge link for UE k is then given as

$$\tau_{E,k}^{\text{ul}} = \frac{b_{I,k}}{R_k^{\text{ul}}}. \tag{12}$$

Among the received $b_{I,k}$ bits from UE $k \in \mathcal{N}_{U,i}$, EN i processes only $c_k b_{I,k}$ bits using its edge server and forwards the remaining $(1-c_k)b_{I,k}$ bits to the CP on the fronthaul link for cloud computing. We denote the partial capacity of the fronthaul link between EN i and CP that is used for transferring the $(1-c_k)b_{I,k}$ input bits for UE k by $C_{F,k}^{\text{ul}} \geq 0$ so that $C_{F,k}^{\text{ul}}$, $k \in \mathcal{N}_{U,i}$, satisfy the constraint

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} C_{F,k}^{\text{ul}} \leq C_F^{\text{ul}}, \tag{13}$$

for all $i \in \mathcal{N}_E$. For given $C_{F,k}^{\text{ul}}$, the uplink fronthaul latency $\tau_{F,k}^{\text{ul}}$ of UE k is given as

$$\tau_{F,k}^{\text{ul}} = \frac{(1-c_k)b_{I,k}}{C_{F,k}^{\text{ul}}}. \tag{14}$$

The CP processes the received $(1-c_k)b_{I,k}$ bits for UE k producing output information of $(1-c_k)b_{O,k}$ bits. The output bits are transmitted to EN i_k that serves UE k. We denote by $C_{F,k}^{\text{dl}} \geq 0$ the partial capacity of the fronthaul link from CP to EN i_k that is used to transfer the $(1-c_k)b_{O,k}$ bits for UE k. Thus, the following constraint should be satisfied:

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} C_{F,k}^{\text{dl}} \leq C_F^{\text{dl}}, \tag{15}$$

for all $i \in \mathcal{N}_E$. The downlink fronthaul latency $\tau_{F,k}^{\text{dl}}$ of UE k for given C_F^{dl} is given as

$$\tau_{F,k}^{\text{dl}} = \frac{(1-c_k)b_{O,k}}{C_{F,k}^{\text{dl}}}. \tag{16}$$

In the downlink, each EN i reports the computation output information of $b_{O,k}$ bits to UE $k \in \mathcal{N}_{U,i}$. To this end, EN i encodes the output information with Gaussian channel codebook producing an encoded baseband signal $\mathbf{s}_k^{\text{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_k^{\text{dl}})$ with $\mathbb{E}[||\mathbf{x}_k^{\text{dl}}||^2] = \text{tr}(\mathbf{Q}_k^{\text{dl}}) \leq P^{\text{dl}}$. Therefore, EN i transmits the encoded signal $\mathbf{s}_k^{\text{dl}}$ during a fraction u_k^{dl} of the downlink time slot. For given $\mathbf{Q}_k^{\text{dl}}$, the achievable downlink data rate R_k^{dl} is given as $R_k^{\text{dl}} = u_k^{\text{dl}} W^{\text{dl}} I(\mathbf{s}_k^{\text{dl}}; y_k^{\text{dl}})$ with $I(\mathbf{s}_k^{\text{dl}}; y_k^{\text{dl}})$ computed as

$$I\left(\mathbf{s}_k^{\text{dl}}; y_k^{\text{dl}}\right) = \log_2\left(1 + \left(1/\sigma_{z,\text{dl}}^2\right) \mathbf{h}_{k,i}^{\text{dl}H} \mathbf{Q}_k^{\text{dl}} \mathbf{h}_{k,i}^{\text{dl}}\right). \tag{17}$$

The optimal covariance matrix $\mathbf{Q}_k^{\text{dl}\star}$, that maximizes the mutual information in (17) while satisfying the constraint $\text{tr}(\mathbf{Q}_k^{\text{dl}}) \leq P^{\text{dl}}$, implements conjugate beamforming [29] and is given as

$$\mathbf{Q}_k^{\text{dl}\star} = P^{\text{dl}} \tilde{\mathbf{h}}_{k,i}^{\text{dl}} \tilde{\mathbf{h}}_{k,i}^{\text{dl}H}, \tag{18}$$

where $\tilde{\mathbf{h}}_{k,i}^{\text{dl}} = \mathbf{h}_{k,i}^{\text{dl}}/||\mathbf{h}_{k,i}^{\text{dl}}||$. By substituting (18) into (17), we obtain the maximized mutual information value $I(\mathbf{s}_k^{\text{dl}}; y_k^{\text{dl}})$ as

$$I\left(\mathbf{s}_k^{\text{dl}}; y_k^{\text{dl}}\right) = \log_2\left(1 + \left(P^{\text{dl}}/\sigma_{z,\text{dl}}^2\right) \left\|\mathbf{h}_{k,i}^{\text{dl}}\right\|^2\right). \tag{19}$$

The downlink latency $\tau_{E,k}^{\text{dl}}$ for UE k on the wireless edge link is hence given as

$$\tau_{E,k}^{\text{dl}} = \frac{b_{O,k}}{R_k^{\text{dl}}}. \tag{20}$$

Finally, the overall latency $\tau_{T,k}$ for each UE k is given as

$$\tau_{T,k} = \tau_{E,k}^{\text{ul}} + \max\left\{\tau_{E,i_k,k}^{\text{exe}}, \tau_{F,k}^{\text{ul}} + \tau_{C,k}^{\text{exe}} + \tau_{F,k}^{\text{dl}}\right\} + \tau_{E,k}^{\text{dl}}, \tag{21}$$

where the second term indicates that local edge computing at EN i_k and fronthaul transmissions can take place simultaneously. As a result, the total latency required for completing the tasks of all the participating UEs is given as

$$\tau_T = \max_{k\in\mathcal{N}_U} \tau_{T,k}. \tag{22}$$

We tackle the problem of optimizing the variables $\mathbf{c} \triangleq \{c_k\}_{k\in\mathcal{N}_U}$, $\mathbf{u}$, $\mathbf{F} \triangleq \{F_{E,i,k}\}_{i\in\mathcal{N}_E, k\in\mathcal{N}_{U,i}} \cup \{F_{C,k}\}_{k\in\mathcal{N}_U}$ and $\mathbf{C}_F \triangleq \{C_{F,k}^{\text{ul}}, C_{F,k}^{\text{dl}}\}_{k\in\mathcal{N}_U}$ with the goal of minimizing the total latency τ_T. We formulate this problem as

$$\underset{\substack{\mathbf{c}\geq 0, \mathbf{u}\geq 0, \mathbf{F}\geq 0,\\ \mathbf{C}_F\geq 0, \boldsymbol{\tau}}}{\text{minimize}} \quad \max_{k\in\mathcal{N}_U} \tau_{T,k} \tag{23a}$$

$$\text{s.t.} \quad \tau_{E,k}^{\text{ul}} \geq \frac{b_{I,k}}{u_k^{\text{ul}} \tilde{R}_k^{\text{ul}}},\ k \in \mathcal{N}_U, \tag{23b}$$

$$\tau_{F,k}^{\text{ul}} \geq \frac{(1-c_k)b_{I,k}}{C_{F,k}^{\text{ul}}},\ k \in \mathcal{N}_U, \tag{23c}$$

$$\tau_{E,k}^{\text{dl}} \geq \frac{b_{O,k}}{u_k^{\text{dl}} \tilde{R}_k^{\text{dl}}},\ k \in \mathcal{N}_U, \tag{23d}$$

$$\tau_{F,k}^{\text{dl}} \geq \frac{(1-c_k)b_{O,k}}{C_{F,k}^{\text{dl}}}\ k \in \mathcal{N}_U, \tag{23e}$$

$$\tau_{E,i_k,k}^{\text{exe}} \geq \frac{c_k b_{I,k} V_k}{F_{E,i_k,k}},\ k \in \mathcal{N}_U, \tag{23f}$$

$$\tau_{C,k}^{\text{exe}} \geq \frac{(1-c_k)b_{I,k} V_k}{F_{C,k}},\ k \in \mathcal{N}_U, \tag{23g}$$

$$c_k \in [0,1],\ k \in \mathcal{N}_U, \tag{23h}$$

$$\sum\nolimits_{k\in\mathcal{N}_U} u_k^{\text{ul}} = \sum\nolimits_{k\in\mathcal{N}_U} u_k^{\text{dl}} = 1, \tag{23i}$$

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} F_{E,i,k} \leq F_{E,i},\ i \in \mathcal{N}_E, \tag{23j}$$

$$\sum\nolimits_{k\in\mathcal{N}_U} F_{C,k} \leq F_C, \tag{23k}$$

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} C_{F,k}^{\text{ul}} \leq C_F^{\text{ul}},\ i \in \mathcal{N}_E, \tag{23l}$$

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} C_{F,k}^{\text{dl}} \leq C_F^{\text{dl}},\ i \in \mathcal{N}_E, \tag{23m}$$

with the notations $\tilde{R}_k^{\text{ul}} = W^{\text{ul}} I(x_k^{\text{ul}}; \mathbf{y}_{i_k}^{\text{ul}})$, $\tilde{R}_k^{\text{dl}} = W^{\text{dl}} I(\mathbf{s}_k^{\text{dl}}; y_k^{\text{dl}})$, and $\boldsymbol{\tau} = \{\tau_{E,k}^{\text{ul}}, \tau_{F,k}^{\text{ul}}, \tau_{E,k}^{\text{dl}}, \tau_{F,k}^{\text{dl}}, \tau_{E,i_k,k}^{\text{exe}}, \tau_{C,k}^{\text{exe}}\}_{k\in\mathcal{N}_U}$.

The problem (23) is non-convex due to the constraints (23c) and (23e)-(23g). We can tackle the non-convex problem by coordinate descent approach [30, Sec. 1.8], since the problem becomes convex if we fix one of the variable sets $\mathbf{c}$ and $\{\mathbf{F}, \mathbf{C}_F\}$. However, the coordinate descent approach cannot

be directly applied to the problems that will be discussed in Sec. III-B and IV, and hence we consider FP [27] as a solution method, which can overcome this limitation.

We observe that all the constraints (23c) and (23e)-(23g), that induce the non-convexity of the problem (23), can be expressed as a function of ratios of optimization variables. It was shown in [27] that FP is suitable for approximating those constraints by convex constraints. In more detail, based on [27, Cor. 1], we can show that, for any real values $\lambda^{\mathrm{ul}}_{F,k}$, $\lambda^{\mathrm{dl}}_{F,k}$, $\lambda^{\mathrm{exe}}_{E,i_k,k}$ and $\lambda^{\mathrm{exe}}_{C,k}$, the following constraints are stricter than (23c) and (23e)-(23g):

$$2\lambda^{\mathrm{ul}}_{F,k}\sqrt{\tau^{\mathrm{ul}}_{F,k}} - (\lambda^{\mathrm{ul}}_{F,k})^2\,(1-c_k) \geq \frac{b_{I,k}}{C^{\mathrm{ul}}_{F,k}},\ k\in\mathcal{N}_U, \quad (24a)$$

$$2\lambda^{\mathrm{dl}}_{F,k}\sqrt{\tau^{\mathrm{dl}}_{F,k}} - (\lambda^{\mathrm{dl}}_{F,k})^2\,(1-c_k) \geq \frac{b_{O,k}}{C^{\mathrm{dl}}_{F,k}},\ k\in\mathcal{N}_U, \quad (24b)$$

$$2\lambda^{\mathrm{exe}}_{E,i_k,k}\sqrt{\tau^{\mathrm{exe}}_{E,i_k,k}} - (\lambda^{\mathrm{exe}}_{E,i_k,k})^2 c_k \geq \frac{b_{I,k}V_k}{F_{E,i_k,k}}, k\in\mathcal{N}_U, \quad (24c)$$

$$2\lambda^{\mathrm{exe}}_{C,k}\sqrt{\tau^{\mathrm{exe}}_{C,k}} - (\lambda^{\mathrm{exe}}_{C,k})^2\,(1-c_k) \geq \frac{b_{I,k}V_k}{F_{C,k}},\ k\in\mathcal{N}_U. \quad (24d)$$

The above constraints have the following desirable properties: they are convex constraints, if the auxiliary variables $\lambda^{\mathrm{ul}}_{F,k}$, $\lambda^{\mathrm{dl}}_{F,k}$, $\lambda^{\mathrm{exe}}_{E,i_k,k}$ and $\lambda^{\mathrm{exe}}_{C,k}$ are fixed. And they become equivalent to (23c) and (23e)-(23g), if the variables $\lambda^{\mathrm{ul}}_{F,k}$, $\lambda^{\mathrm{dl}}_{F,k}$, $\lambda^{\mathrm{exe}}_{E,i_k,k}$ and $\lambda^{\mathrm{exe}}_{C,k}$ are given as

$$\lambda^{\mathrm{ul}}_{F,k} = \frac{\sqrt{\tau^{\mathrm{ul}}_{F,k}}}{1-c_k},\ \lambda^{\mathrm{dl}}_{F,k} = \frac{\sqrt{\tau^{\mathrm{dl}}_{F,k}}}{1-c_k},\ \lambda^{\mathrm{exe}}_{E,i_k,k} = \frac{\sqrt{\tau^{\mathrm{exe}}_{E,i_k,k}}}{c_k},$$
$$\text{and } \lambda^{\mathrm{exe}}_{C,k} = \frac{\sqrt{\tau^{\mathrm{exe}}_{C,k}}}{1-c_k}. \quad (25)$$

Based on the above observation, we consider the problem obtained by replacing the constraints (23c) and (23e)-(23g) with (24) in (23) and adding $\boldsymbol{\lambda} = \{\lambda^{\mathrm{ul}}_{F,k}, \lambda^{\mathrm{dl}}_{F,k}, \lambda^{\mathrm{exe}}_{E,i_k,k}, \lambda^{\mathrm{exe}}_{C,k}\}_{k\in\mathcal{N}_U}$ as optimization variables. To tackle the obtained problem, which has the same optimal value as (23), we propose an iterative algorithm, in which the variables $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$ and $\boldsymbol{\lambda}$ are alternately updated. Since the optimization of $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$ for fixed $\boldsymbol{\lambda}$ is a convex problem, standard convex solvers, such as the CVX software [31], can be used. The optimal $\boldsymbol{\lambda}$ for fixed $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$ can be obtained as (25), which make the constraints (24a)-(24d) equivalent to the original constraints (23c) and (23e)-(23g). We describe the detailed algorithm in Algorithm 1.

The convex problem solved at Step 4 of each tth iteration in Algorithm 1 has stricter constraints than the original problem (23). Also, the feasible space of the convex problem contains the solution obtained at the $(t-1)$th iteration. Thus, the solution of the convex problem at the tth iteration belongs to the feasible space of problem (23) and achieves a lower latency value than the solution of the $(t-1)$th iteration. Therefore, Algorithm 1 produces monotonically decreasing latency values with respect to the iteration index t so that it converges to a locally optimal point. For more formal proof of the convergence of SCA and FP algorithms, we refer to [11], [27]. We can operate Algorithm 1 with an arbitrary initial point that satisfies the conditions (23b)-(23m). In the simulation section, we initialize the variables $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F\}$ at Step 1 as

$$u_k \leftarrow 1/N_U,\ k\in\mathcal{N}_U, \quad (26a)$$
$$c_k \leftarrow 1/2,\ k\in\mathcal{N}_U, \quad (26b)$$
$$F_{E,i,k} \leftarrow F_{E,i}/|\mathcal{N}_{U,i}|,\ k\in\mathcal{N}_{U,i}, i\in\mathcal{N}_E, \quad (26c)$$
$$F_{C,k} \leftarrow F_C/N_U,\ k\in\mathcal{N}_U, \quad (26d)$$
$$C^m_{F,k} \leftarrow C^{\mathrm{ul}}_F/|\mathcal{N}_{U,i}|,\ k\in\mathcal{N}_{U,i}, i\in\mathcal{N}_E, m\in\{\mathrm{ul},\mathrm{dl}\}. \quad (26e)$$

For the given $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F\}$, we compute an initial value for $\boldsymbol{\tau}$ according to (12), (14), (16), and (20).

Algorithm 1 Alternating optimization algorithm that tackles problem (23)

1. Initialize $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$ as arbitrary values that satisfy the constraints (23b)-(23m), and set $t\leftarrow 1$.
2. Calculate the total latency τ_T in (22) with the initialized $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$, and set $\tau_T^{(0)} \leftarrow \tau_T$.
3. Set $\boldsymbol{\lambda}$ according to (25).
4. Update the variables $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$ as a solution of the convex problem which is obtained by replacing the constraints (23c) and (23e)-(23g) with (24a)-(24d) and then by fixing $\boldsymbol{\lambda}$.
5. Calculate the total latency τ_T with the updated $\{\mathbf{c},\mathbf{u},\mathbf{F},\mathbf{C}_F,\boldsymbol{\tau}\}$, and set $\tau_T^{(t)} \leftarrow \tau_T$.
6. Stop if $|\tau_T^{(t)} - \tau_T^{(t-1)}| \leq \delta$ or $t > t_{\max}$. Otherwise, set $t\leftarrow t+1$ and go back to Step 2.

The complexity of Algorithm 1 is given by the number of iterations multiplied by the complexity of solving the convex problem at each iteration (i.e., Step 4). The complexity of solving a generic convex problem is upper bounded by $\mathcal{O}(n(n^3+M)\log(1/\epsilon))$ [32, p. 4], where n denotes the number of optimization variables, M is the number of arithmetic operations required to compute the objective and constraint functions, and ϵ represents the desired error tolerance. The numbers n and M equal $n = 13N_U$ and $M = 45N_U$, respectively, for the convex problem solved at Step 4 of Algorithm 1. However, to the best of our knowledge, the analysis of the convergence rate of general SCA algorithms is still an open problem. Instead, we provide some numerical evidence of the fast convergence of Algorithm 1 in Sec. V.

B. Non-Orthogonal Multiple Access

In this subsection, we discuss the design with non-orthogonal multiple access. With non-orthogonal access, N_U UEs communicate simultaneously with N_E ENs on the same time and frequency resource. Therefore, the uplink and downlink communications on the wireless edge link are impaired by inter-UE interference signals, while benefiting from transmission on a larger time interval. The computation and fronthaul transmission models are the same as the one described in Sec. III-A, and we detail here only the uplink and downlink communication phases and the resulting latency performance.

As in Sec. III-A, we assume that each UE k uses a Gaussian channel codebook so that its transmitted signal x^{ul}_k is distributed as $x^{\mathrm{ul}}_k \sim \mathcal{CN}(0, p^{\mathrm{ul}}_k)$. The transmit power p^{ul}_k is

subject to the constraint $p_k^{\mathrm{ul}} \in [0, P^{\mathrm{ul}}]$. Due to the presence of inter-UE interference signals, full power transmission at all UEs may cause an optimality loss. This suggests that we need to carefully design the transmit power variables p_k^{ul}, $k \in \mathcal{N}_U$, by adapting to channel state information (CSI).

Each EN i needs to decode the signals $\{x_k^{\mathrm{ul}}\}_{k\in\mathcal{N}_{U,i}}$ based on the received signal $\mathbf{y}_i^{\mathrm{ul}}$. We assume that the signals $\{x_k^{\mathrm{ul}}\}_{k\in\mathcal{N}_{U,i}}$ are detected in parallel without successive interference cancellation (SIC) as in [33], [34] in order to minimize the decoding delay. We leave the design and analysis with SIC decoding [35] while taking into account the decoding delay for future work.

Under the assumption of parallel decoding, the achievable rate R_k^{ul} of UE k in the uplink channel is given as $R_k^{\mathrm{ul}} = W^{\mathrm{ul}} I(x_k^{\mathrm{ul}}; \mathbf{y}_{i_k}^{\mathrm{ul}})$ with the mutual information value computed as

$$I(x_k^{\mathrm{ul}}; \mathbf{y}_{i_k}^{\mathrm{ul}}) = f_{E,k}^{\mathrm{ul}}(\mathbf{p}) = \Psi\left(p_k^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}\,H}, \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \sum\nolimits_{l\in\mathcal{N}_U\setminus\{k\}} p_l^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}\,H}\right). \tag{27}$$

Here we have defined the notation $\mathbf{p} \triangleq \{p_k^{\mathrm{ul}}\}_{k\in\mathcal{N}_U}$, and the function

$$\Psi(\mathbf{A}, \mathbf{B}) = \log_2 \det\left(\mathbf{I} + \mathbf{B}^{-1}\mathbf{A}\right) \tag{28}$$

For given R_k^{ul}, the uplink edge latency $\tau_{E,k}^{\mathrm{ul}}$ for UE k is given as (12).

For the downlink edge link, each EN i transmits a superposition of the signals $\mathbf{s}_k^{\mathrm{dl}}$, $k \in \mathcal{N}_{U,i}$, where $\mathbf{s}_k^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_k^{\mathrm{dl}})$ encodes the task output of UE k. The transmit signal of EN i is written as

$$\mathbf{x}_i^{\mathrm{dl}} = \sum\nolimits_{k\in\mathcal{N}_{U,i}} \mathbf{s}_k^{\mathrm{dl}}. \tag{29}$$

With the above transmission model, the downlink transmit power constraint (9) can be expressed as $\sum_{k\in\mathcal{N}_{U,i}} \mathrm{tr}(\mathbf{Q}_k^{\mathrm{dl}}) \leq P^{\mathrm{dl}}$, and the achievable rate R_k^{dl} of UE k on the wireless edge link is given as $R_k^{\mathrm{dl}} = W^{\mathrm{dl}} I(\mathbf{s}_k^{\mathrm{dl}}; y_k^{\mathrm{dl}})$ with

$$I\left(\mathbf{s}_k^{\mathrm{dl}}; y_k^{\mathrm{dl}}\right) = f_{E,k}^{\mathrm{dl}}(\mathbf{Q}) = \Psi\left(\mathbf{h}_{k,i_k}^{\mathrm{dl}\,H} \mathbf{Q}_k^{\mathrm{dl}} \mathbf{h}_{k,i_k}^{\mathrm{dl}}, \sigma_{z,\mathrm{dl}}^2 + \sum\nolimits_{l\in\mathcal{N}_U\setminus\{k\}} \mathbf{h}_{k,i_l}^{\mathrm{dl}\,H} \mathbf{Q}_l^{\mathrm{dl}} \mathbf{h}_{k,i_l}^{\mathrm{dl}}\right), \tag{30}$$

where $\mathbf{Q} \triangleq \{\mathbf{Q}_k^{\mathrm{dl}}\}_{k\in\mathcal{N}_U}$. For given R_k^{dl}, the downlink edge latency $\tau_{E,k}^{\mathrm{dl}}$ of UE k is given as (20).

For the non-orthogonal multiple access scheme as described above, we aim at jointly optimizing the variables $\mathbf{p}$, $\mathbf{Q}$, $\mathbf{c}$, $\mathbf{F}$ and $\mathbf{C}_F$ with the goal of minimizing the total latency τ_T in (22). The problem can be written as

$$\underset{\substack{\mathbf{p}\geq\mathbf{0},\mathbf{Q}\succeq\mathbf{0},\mathbf{c}\geq\mathbf{0},\\ \mathbf{F}\geq\mathbf{0},\mathbf{C}_F\geq\mathbf{0},\boldsymbol{\tau},\mathbf{R}}}{\text{minimize}} \ \max_{k\in\mathcal{N}_\mathcal{U}} \tau_{T,k} \tag{31a}$$

$$\text{s.t.} \quad \tau_{E,k}^{\mathrm{ul}} \geq \frac{b_{I,k}}{R_k^{\mathrm{ul}}},\ k \in \mathcal{N}_U, \tag{31b}$$

$$\tau_{E,k}^{\mathrm{dl}} \geq \frac{b_{O,k}}{R_k^{\mathrm{dl}}},\ k \in \mathcal{N}_U, \tag{31c}$$

$$\text{(23c), (23e)-(23g)}, \tag{31d}$$

$$R_{E,k}^{\mathrm{ul}} \leq f_{E,k}^{\mathrm{ul}}(\mathbf{p}),\ k \in \mathcal{N}_U, \tag{31e}$$

$$R_{E,k}^{\mathrm{dl}} \leq f_{E,k}^{\mathrm{dl}}(\mathbf{Q}),\ k \in \mathcal{N}_U, \tag{31f}$$

$$p_k^{\mathrm{ul}} \leq P^{\mathrm{ul}},\ k \in \mathcal{N}_U, \tag{31g}$$

$$\sum\nolimits_{k\in\mathcal{N}_{U,i}} \mathrm{tr}\left(\mathbf{Q}_k^{\mathrm{dl}}\right) \leq P^{\mathrm{dl}},\ i \in \mathcal{N}_E, \tag{31h}$$

$$c_k \in [0,1],\ k \in \mathcal{N}_U, \tag{31i}$$

$$\text{(23j)-(23m)}, \tag{31j}$$

where we have defined $\mathbf{R} \triangleq \{R_{E,k}^{\mathrm{ul}}, R_{E,k}^{\mathrm{dl}}\}_{k\in\mathcal{N}_U}$.

We note that it is more challenging to tackle problem (31) than (23) due to the presence of inter-UE interference signals on the wireless edge links. Accordingly, the uplink and downlink transmission strategies on edge links, which are characterized by the variables $\mathbf{p}$ and $\mathbf{Q}$, need to be jointly optimized. Also, the constraints (31e) and (31f) on the edge throughputs, which involve matrix variables $\mathbf{Q}$, are not convex. To address these complications, we employ FP [27] as well as matrix FP [28], which is a generalized version of [27].

We first observe that the constraints (31d), that are expressed as a function of ratios of scalar optimization variables, can be handled by FP [27] as in Sec. III-A. Based on [27, Cor. 1], we replace the constraints (31d) with stricter constraints (24a)-(24d), which become equivalent to (31d) if the variables $\lambda_{F,k}^{\mathrm{ul}}$, $\lambda_{F,k}^{\mathrm{dl}}$, $\lambda_{E,i_k,k}^{\mathrm{exe}}$ and $\lambda_{C,k}^{\mathrm{exe}}$ equal (25).

The other non-convex constraints (31e) and (31f) contain ratios of matrix variables. Thus, we need to employ matrix FP [28], which generalizes scalar or vector version of FP in [27]. From [28, Cor. 1], the following constraints are stricter than (31e) and (31f) for any $\Gamma_{E,k}^{\mathrm{ul}} \in \mathbb{C}^{1\times 1}$, $\boldsymbol{\theta}_{E,k}^{\mathrm{ul}} \in \mathbb{C}^{n_{E,i_k}\times 1}$, $\boldsymbol{\Gamma}_{E,k}^{\mathrm{dl}} \in \mathbb{C}^{n_{E,i_k}\times n_{E,i_k}}$ and $\boldsymbol{\theta}_{E,k}^{\mathrm{dl}} \in \mathbb{C}^{1\times n_{E,i_k}}$:

$$R_{E,k}^{\mathrm{ul}} \leq \phi\left(\begin{array}{c}\Gamma_{E,k}^{\mathrm{ul}}, \boldsymbol{\theta}_{E,k}^{\mathrm{ul}}, \tilde{p}_k^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}}, \\ \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \sum_{l\in\mathcal{N}_U} p_l^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}\,H}\end{array}\right), \text{ and} \tag{32a}$$

$$R_{E,k}^{\mathrm{dl}} \leq \phi\left(\begin{array}{c}\boldsymbol{\Gamma}_{E,k}^{\mathrm{dl}}, \boldsymbol{\theta}_{E,k}^{\mathrm{dl}}, \mathbf{h}_{k,i_k}^{\mathrm{dl}\,H} \tilde{\mathbf{Q}}_{E,k}^{\mathrm{dl}}, \\ \sigma_{z,\mathrm{dl}}^2 + \sum_{l\in\mathcal{N}_U} \mathbf{h}_{k,i_l}^{\mathrm{dl}\,H} \mathbf{Q}_{E,l}^{\mathrm{dl}} \mathbf{h}_{k,i_l}^{\mathrm{dl}}\end{array}\right), \tag{32b}$$

where we have defined the variables $\tilde{p}_k^{\mathrm{ul}} = \sqrt{p_k^{\mathrm{ul}}} \in [0, \sqrt{P^{\mathrm{ul}}}]$, $\tilde{\mathbf{Q}}_{E,k}^{\mathrm{dl}} = \mathbf{Q}_{E,k}^{\mathrm{dl}\,1/2}$, and the function

$$\phi(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) = \log_2 \det(\mathbf{I} + \mathbf{A}) - \frac{1}{\ln 2}\mathrm{tr}(\mathbf{A}) + \frac{1}{\ln 2}\mathrm{tr}\left((\mathbf{I} + \mathbf{A})\left(2\mathbf{C}^H\mathbf{B} - \mathbf{B}^H\mathbf{D}\mathbf{B}\right)\right). \tag{33}$$

Also, the above constraints are equivalent to (31e) and (31f) if

$$\Gamma_{E,k}^{\mathrm{ul}} = p_k^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}\,H} \left(\sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \sum_{l \in \mathcal{N}_U \setminus \{k\}} p_l^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}\,H} \right)^{-1} \mathbf{h}_{i_k,k}^{\mathrm{ul}}, \quad (34a)$$

$$\boldsymbol{\theta}_{E,k}^{\mathrm{ul}} = \tilde{p}_k^{\mathrm{ul}} \left(\sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \sum_{l \in \mathcal{N}_U} p_l^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}\,H} \right)^{-1} \mathbf{h}_{i_k,k}^{\mathrm{ul}}, \quad (34b)$$

$$\boldsymbol{\Gamma}_{E,k}^{\mathrm{dl}} = \tilde{\mathbf{Q}}_{E,k}^{\mathrm{dl}\,H} \mathbf{h}_{k,i_k}^{\mathrm{dl}} \left(\sigma_{z,\mathrm{dl}}^2 + \sum_{l \in \mathcal{N}_U \setminus \{k\}} \mathbf{h}_{k,i_l}^{\mathrm{dl}\,H} \mathbf{Q}_{E,l}^{\mathrm{dl}} \mathbf{h}_{k,i_l}^{\mathrm{dl}} \right)^{-1} \times \mathbf{h}_{k,i_k}^{\mathrm{dl}\,H} \tilde{\mathbf{Q}}_{E,k}^{\mathrm{dl}}, \text{ and} \quad (34c)$$

$$\boldsymbol{\theta}_{E,k}^{\mathrm{dl}} = \left(\sigma_{z,\mathrm{dl}}^2 + \sum_{l \in \mathcal{N}_U} \mathbf{h}_{k,i_l}^{\mathrm{dl}\,H} \mathbf{Q}_{E,l}^{\mathrm{dl}} \mathbf{h}_{k,i_l}^{\mathrm{dl}} \right)^{-1} \mathbf{h}_{k,i_k}^{\mathrm{dl}\,H} \tilde{\mathbf{Q}}_{E,k}^{\mathrm{dl}}. \quad (34d)$$

Using the alternative representations (24) and (32) to the non-convex constraints (31d)-(31f), we restate the problem (31) with additional optimization variables $\boldsymbol{\lambda}$, $\boldsymbol{\Gamma} \triangleq \{\Gamma_{E,k}^{\mathrm{ul}}, \boldsymbol{\Gamma}_{E,k}^{\mathrm{dl}}\}_{k \in \mathcal{N}_U}$ and $\boldsymbol{\theta} \triangleq \{\boldsymbol{\theta}_{E,k}^{\mathrm{ul}}, \boldsymbol{\theta}_{E,k}^{\mathrm{dl}}\}_{k \in \mathcal{N}_U}$. We tackle the obtained problem by alternately optimizing the variables $\{\mathbf{c}, \tilde{\mathbf{p}}, \tilde{\mathbf{Q}}, \mathbf{F}, \mathbf{C}_F, \boldsymbol{\tau}, \mathbf{R}\}$ and $\{\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\theta}\}$. The detailed algorithm is summarized in Algorithm 2. Similarly to Algorithm 1, Algorithm 2 achieves monotonically decreasing latency with respect to the number of iterations, whose solution converges to a locally optimal point of (31) due to its non-convexity. In Sec. V, we initialize the variables $\{\mathbf{c}, \mathbf{F}, \mathbf{C}_F\}$ and $\{\tilde{\mathbf{p}}, \tilde{\mathbf{Q}}\}$ as (26b)-(26e) and

$$\tilde{p}_k \leftarrow \sqrt{P^{\mathrm{ul}}},\ k \in \mathcal{N}_U, \quad (35a)$$

$$\tilde{\mathbf{Q}}_k^{\mathrm{dl}} \leftarrow \sqrt{\frac{P^{\mathrm{dl}}}{\sum_{l \in \mathcal{N}_{U,i}} ||\mathbf{V}_l^{\mathrm{dl}}||_F^2}} \mathbf{V}_k^{\mathrm{dl}},\ k \in \mathcal{N}_{U,i}, i \in \mathcal{N}_E, \quad (35b)$$

respectively, where the elements of $\mathbf{V}_k^{\mathrm{dl}} \in \mathbb{C}^{n_{E,i} \times n_{E,i}}$, $k \in \mathcal{N}_{U,i}$, are independent and identically distributed as $\mathcal{CN}(0,1)$. For the given $\{\mathbf{c}, \mathbf{F}, \mathbf{C}_F, \tilde{\mathbf{p}}, \tilde{\mathbf{Q}}\}$, we compute the rates $\mathbf{R}$ using (27) and (30), from which the latency variables $\boldsymbol{\tau}$ can be initialized as (12), (14), (16), and (20).

The complexity of Algorithm 2 is given as the product of the number of iterations and the complexity of solving the convex problem at Step 4. The complexity of the latter is upper bounded by $\mathcal{O}(n(n^3 + M)\log(1/\epsilon))$ [32, p. 4], where the numbers of optimization variables and arithmetic operations are given as $n = N_U(4\tilde{n}_E^2 + 14)$ and $M = N_U(\tilde{n}_E(14\tilde{n}_E + 1) + 41) + \tilde{n}_E(8\tilde{n}_E^2 + 5\tilde{n}_E + 3)$, respectively. Here we have assumed that every EN uses the same number $\tilde{n}_E$ of antennas, i.e., $n_{E,i} = \tilde{n}_E$ for all $i \in \mathcal{N}_E$. Some numerical evidence of the convergence rate of Algorithm 2 is provided in Sec. V.

Algorithm 2 Alternating optimization algorithm that tackles problem (31)

1. Initialize $\{\mathbf{c}, \tilde{\mathbf{p}}, \tilde{\mathbf{Q}}, \mathbf{F}, \mathbf{C}_F, \boldsymbol{\tau}, \mathbf{R}\}$ as arbitrary values/matrices that satisfy the constraints (31b)-(31j), and set $t \leftarrow 1$.
2. Calculate the total latency τ_T in (22) with the initialized $\{\mathbf{c}, \tilde{\mathbf{p}}, \tilde{\mathbf{Q}}, \mathbf{F}, \mathbf{C}_F, \boldsymbol{\tau}, \mathbf{R}\}$, and set $\tau_T^{(0)} \leftarrow \tau_T$.
3. Set $\{\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\theta}\}$ according to (25) and (34).
4. Update the variables $\{\mathbf{c}, \tilde{\mathbf{p}}, \tilde{\mathbf{Q}}, \mathbf{F}, \mathbf{C}_F, \boldsymbol{\tau}, \mathbf{R}\}$ as a solution of the convex problem which is obtained by replacing the constraints (31d)-(31f) with (24a)-(24d) in (23), (32a) and (32b) and then by fixing $\{\boldsymbol{\lambda}, \boldsymbol{\Gamma}, \boldsymbol{\theta}\}$.
5. Calculate the total latency τ_T in (22) with the updated $\{\mathbf{c}, \tilde{\mathbf{p}}, \tilde{\mathbf{Q}}, \mathbf{F}, \mathbf{C}_F, \boldsymbol{\tau}, \mathbf{R}\}$, and set $\tau_T^{(t)} \leftarrow \tau_T$.
6. Stop if $|\tau_T^{(t)} - \tau_T^{(t-1)}| \leq \delta$ or $t > t_{\max}$. Otherwise, set $t \leftarrow t + 1$ and go back to Step 2.

IV. Optimization for the C-RAN Architecture

In this section, we investigate the design of collaborative cloud and edge mobile computing system within a C-RAN architecture [18]–[20]. In C-RAN, the baseband signals of distributed ENs are processed by the CP in a centralized manner for the purpose of effective interference management. In the following subsections, we describe the uplink and downlink communication phases and the total end-to-end latency required for completing all the tasks, and discuss the joint optimization of C-RAN signal processing and computational resource allocation strategies.

A. Uplink Communication and Latency

As illustrated in Sec. II-A, each UE k splits its computation input information into two parts of $c_k b_{I,k}$ and $(1-c_k)b_{I,k}$ bits, and sends the former and latter parts to its serving EN i_k and the CP, respectively. In the D-RAN protocol detailed in Sec. III, both parts were encoded into a single codeword, since all the input information had to be decoded by the serving EN i_k. However, in the C-RAN scheme, only one part is decoded by EN i_k, and the other codeword is decoded by the CP based on the fronthaul received signals. To accommodate this requirement, we leverage superposition coding as discussed next.

We denote the encoded signals for the two parts of $c_k b_{I,k}$ and $(1 - c_k)b_{I,k}$ bits by $s_{E,k}^{\mathrm{ul}}$ and $s_{C,k}^{\mathrm{ul}}$, respectively. Under independent Gaussian channel codebooks, the two signals are independent of each other and distributed as $s_{E,k}^{\mathrm{ul}} \sim \mathcal{CN}(0, p_{E,k}^{\mathrm{ul}})$ and $s_{C,k}^{\mathrm{ul}} \sim \mathcal{CN}(0, p_{C,k}^{\mathrm{ul}})$. UE k transmits a superposition of the encoded signals so that the transmit signal x_k^{ul} is given as

$$x_k^{\mathrm{ul}} = s_{E,k}^{\mathrm{ul}} + s_{C,k}^{\mathrm{ul}}, \quad (36)$$

and the transmit power constraint (8) can be written as $p_{E,k}^{\mathrm{ul}} + p_{C,k}^{\mathrm{ul}} \leq P^{\mathrm{ul}}$.

Based on the uplink received signal $\mathbf{y}_i^{\mathrm{ul}}$, EN i detects the signals $s_{E,k}^{\mathrm{ul}}$ transmitted by its serving UEs $k \in \mathcal{N}_{U,i}$. The achievable rate $R_{E,k}^{\mathrm{ul}}$ of each signal $s_{E,k}^{\mathrm{ul}}$ in bps is given as $R_{E,k}^{\mathrm{ul}} = W^{\mathrm{ul}} I(s_{E,k}^{\mathrm{ul}}; \mathbf{y}_{i_k}^{\mathrm{ul}})$ with

$$I\left(s_{E,k}^{\mathrm{ul}}; \mathbf{y}_{i_k}^{\mathrm{ul}}\right) = f_{E,k}^{\mathrm{ul}}\left(\mathbf{p}^{\mathrm{ul}}\right) = \quad (37)$$
$$\Psi\left(p_{E,k}^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}H}, \left(\begin{array}{c} \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \sum_{l \in \mathcal{N}_U \setminus \{k\}} p_{E,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}\,H} \\ + \sum_{l \in \mathcal{N}_U} p_{C,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}\,H} \end{array} \right)\right).$$

Here we have defined $\mathbf{p}^{\mathrm{ul}} \triangleq \{p_{E,k}^{\mathrm{ul}}, p_{C,k}^{\mathrm{ul}}\}_{k \in \mathcal{N}_U}$.

After the local decoding described above, EN i cancels out the impact of the decoded signals from the received signal $\mathbf{y}_i^{\text{ul}}$ as

$$\tilde{\mathbf{y}}_i^{\text{ul}} \leftarrow \mathbf{y}_i^{\text{ul}} - \sum\nolimits_{l\in\mathcal{N}_{U,i}} \mathbf{h}_{i,l}^{\text{ul}} s_{E,l}^{\text{ul}}. \tag{38}$$

Since the fronthaul link connecting EN i to the CP has finite capacity C_F bps, a quantized version of the signal $\tilde{\mathbf{y}}_i^{\text{ul}}$, denoted by $\hat{\mathbf{y}}_i^{\text{ul}}$, is forwarded to the CP. We assume the Gaussian test channel as in [19], [20]. Then, the quantized signal $\hat{\mathbf{y}}_i^{\text{ul}}$ is modeled as

$$\hat{\mathbf{y}}_i^{\text{ul}} = \tilde{\mathbf{y}}_i^{\text{ul}} + \mathbf{q}_i^{\text{ul}}, \tag{39}$$

where the quantization distortion noise $\mathbf{q}_i^{\text{ul}}$ is independent of $\tilde{\mathbf{y}}_i^{\text{ul}}$ and is distributed as $\mathbf{q}_i^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i^{\text{ul}})$. Under the quantization model (39), the compression rate γ_i^{ul}, that equals the number of bits representing the quantized signal $\hat{\mathbf{y}}_i^{\text{ul}}$ per baseband sample, is given as [36]

$$\begin{aligned}\gamma_i^{\text{ul}} &= I\left(\tilde{\mathbf{y}}_i^{\text{ul}}; \hat{\mathbf{y}}_i^{\text{ul}}\right) = g_i^{\text{ul}}\left(\mathbf{p}^{\text{ul}}, \mathbf{\Omega}_i^{\text{ul}}\right) \\ &= \log_2 \det\left(\begin{array}{c} \sum_{k\in\mathcal{N}_U\setminus\mathcal{N}_{U,i}} p_{E,k}^{\text{ul}} \mathbf{h}_{i,k}^{\text{ul}} \mathbf{h}_{i,k}^{\text{ul}H} + \\ \sum_{k\in\mathcal{N}_U} p_{C,k}^{\text{ul}} \mathbf{h}_{i,k}^{\text{ul}} \mathbf{h}_{i,k}^{\text{ul}H} + \sigma_{z,\text{ul}}^2 \mathbf{I} + \mathbf{\Omega}_i^{\text{ul}} \end{array} \right) \\ &\quad - \log_2 \det\left(\mathbf{\Omega}_i^{\text{ul}}\right). \end{aligned} \tag{40}$$

EN i should send $W^{\text{ul}}\tau_E^{\text{ul}}\gamma_i^{\text{ul}}$ bits to the CP on the fronthaul link of capacity C_F bps, since the duration of each baseband sample is approximately $1/W^{\text{ul}}$ sec, and hence $\tau_E^{\text{ul}}/(1/W^{\text{ul}}) = W^{\text{ul}}\tau_E^{\text{ul}}$ quanzited baseband samples should be forwarded to the CP. Due to the parallel operation of fronthaul links of different ENs, the fronthaul latency τ_F^{ul} for uplink is given as

$$\tau_F^{\text{ul}} = \max_{i\in\mathcal{N}_E} \frac{W^{\text{ul}}\tau_E^{\text{ul}}\gamma_i^{\text{ul}}}{C_F}. \tag{41}$$

The CP recovers the quantized signals $\hat{\mathbf{y}}_1^{\text{ul}}, \hat{\mathbf{y}}_2^{\text{ul}}, \ldots, \hat{\mathbf{y}}_{N_E}^{\text{ul}}$ from the bit streams received on the fronthaul links. The vector $\hat{\mathbf{y}}^{\text{ul}} = [\hat{\mathbf{y}}_1^{\text{ul}H}\, \hat{\mathbf{y}}_2^{\text{ul}H} \cdots \hat{\mathbf{y}}_{N_E}^{\text{ul}H}]^H$, which stacks the quantized signals from all ENs, can be written as

$$\hat{\mathbf{y}}^{\text{ul}} = \sum\nolimits_{l\in\mathcal{N}_U} \tilde{\mathbf{h}}_l^{\text{ul}} s_{E,l}^{\text{ul}} + \sum\nolimits_{l\in\mathcal{N}_U} \mathbf{h}_l^{\text{ul}} s_{C,l}^{\text{ul}} + \mathbf{q}^{\text{ul}} + \mathbf{z}^{\text{ul}}, \tag{42}$$

where we have defined $\mathbf{h}_k^{\text{ul}} = [\mathbf{h}_{1,k}^{\text{ul}H}\, \mathbf{h}_{2,k}^{\text{ul}H} \cdots \mathbf{h}_{N_E,k}^{\text{ul}H}]^H$, $\tilde{\mathbf{h}}_k^{\text{ul}} = [\tilde{\mathbf{h}}_{1,k}^{\text{ul}H}\, \tilde{\mathbf{h}}_{2,k}^{\text{ul}H} \cdots \tilde{\mathbf{h}}_{N_E,k}^{\text{ul}H}]^H$ with $\tilde{\mathbf{h}}_{i,k}^{\text{ul}} = \mathbf{h}_{i,k}\mathbf{1}_{i\neq i_k} + \mathbf{0}_{n_{E,i}\times 1}\mathbf{1}_{i=i_k}$, $\mathbf{q}^{\text{ul}} = [\mathbf{q}_1^{\text{ul}H}\, \mathbf{q}_2^{\text{ul}H} \cdots \mathbf{q}_{N_E}^{\text{ul}H}]^H$, and $\mathbf{z}^{\text{ul}} = [\mathbf{z}_1^{\text{ul}H}\, \mathbf{z}_2^{\text{ul}H} \cdots \mathbf{z}_{N_E}^{\text{ul}H}]^H$. Here $\mathbf{1}_{(\cdot)}$ is an indicator function which takes 1 if the statement in the subscript is true and 0 otherwise. The stacked noise vectors $\mathbf{q}^{\text{ul}}$ and $\mathbf{z}^{\text{ul}}$ are distributed as $\mathbf{q}^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \bar{\mathbf{\Omega}}^{\text{ul}})$ and $\mathbf{z}^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \sigma_{z,\text{ul}}^2\mathbf{I})$, respectively, with $\bar{\mathbf{\Omega}}^{\text{ul}} = \text{diag}(\{\mathbf{\Omega}_i^{\text{ul}}\}_{i\in\mathcal{N}_E})$.

Using the recovered quantized signal vector $\hat{\mathbf{y}}^{\text{ul}}$, the CP detects all the signals $s_{C,k}^{\text{ul}}$, which are necessary for cloud computing. The achievable rate $R_{C,k}^{\text{ul}}$ of the signal $s_{C,k}^{\text{ul}}$ is given as $R_{C,k}^{\text{ul}} = W^{\text{ul}} I(s_{C,k}^{\text{ul}}; \hat{\mathbf{y}}^{\text{ul}})$, where the mutual information value is computed as

$$\begin{aligned} &I\left(s_{C,k}^{\text{ul}}; \hat{\mathbf{y}}^{\text{ul}}\right) = f_{C,k}^{\text{ul}}\left(\mathbf{p}^{\text{ul}}, \mathbf{\Omega}^{\text{ul}}\right) = \\ &\Psi\left(p_{C,k}^{\text{ul}} \mathbf{h}_k^{\text{ul}} \mathbf{h}_k^{\text{ul}H}, \left(\begin{array}{c} \sigma_{z,\text{ul}}^2\mathbf{I} + \sum_{l\in\mathcal{N}_U} p_{E,l}^{\text{ul}} \tilde{\mathbf{h}}_l^{\text{ul}} \tilde{\mathbf{h}}_l^{\text{ul}H} + \\ \sum_{l\in\mathcal{N}_U\setminus\{k\}} p_{C,l}^{\text{ul}} \mathbf{h}_l^{\text{ul}} \mathbf{h}_l^{\text{ul}H} + \bar{\mathbf{\Omega}}^{\text{ul}} \end{array} \right)\right). \end{aligned} \tag{43}$$

Consequently, the latency τ_E^{ul} for uploading the input information of the UEs on the uplink channel is given as

$$\tau_E^{\text{ul}} = \max_{k\in\mathcal{N}_U} \left\{ \frac{c_k b_{I,k}}{W^{\text{ul}} f_{E,k}^{\text{ul}}(\mathbf{p})}, \frac{(1-c_k) b_{I,k}}{W^{\text{ul}} f_{C,k}^{\text{ul}}(\mathbf{p}, \mathbf{\Omega}^{\text{ul}})} \right\}. \tag{44}$$

B. Downlink Communication and Latency

After completing the computation tasks, the CP encodes the computation output information of $(1-c_k)b_{O,k}$ bits for each UE k with Gaussian channel codebook and obtains an encoded baseband signal $\mathbf{s}_{C,k}^{\text{dl}} \in \mathbb{C}^{n_E\times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{C,k}^{\text{dl}})$.

The CP computes a signal vector $\tilde{\mathbf{x}}^{\text{dl}} \in \mathbb{C}^{n_E\times 1}$ by superimposing the encoded signals as

$$\tilde{\mathbf{x}}^{\text{dl}} = \sum\nolimits_{k\in\mathcal{N}_U} \mathbf{s}_{C,k}^{\text{dl}}. \tag{45}$$

The ith subvector $\tilde{\mathbf{x}}_i^{\text{dl}} \in \mathbb{C}^{n_{E,i}\times 1}$ of $\tilde{\mathbf{x}}^{\text{dl}} = [\tilde{\mathbf{x}}_1^{\text{dl}H} \cdots \tilde{\mathbf{x}}_{N_E}^{\text{dl}H}]^H$ is transferred to EN i on the fronthaul link. To this end, it is quantized, and we model the quantized signal $\hat{\mathbf{x}}_i^{\text{dl}}$ under the Gaussian test channel [19], [20] as

$$\hat{\mathbf{x}}_i^{\text{dl}} = \tilde{\mathbf{x}}_i^{\text{dl}} + \mathbf{q}_i^{\text{dl}}, \tag{46}$$

where the quantization distortion noise $\mathbf{q}_i^{\text{dl}}$ is independent of $\tilde{\mathbf{x}}_i^{\text{dl}}$ and distributed as $\mathbf{q}_i^{\text{dl}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_i^{\text{dl}})$.

The compression rate γ_i^{dl} needed for representing the quantized signal $\hat{\mathbf{x}}_i^{\text{dl}}$ in bits per baseband sample is given as

$$\begin{aligned} &\gamma_i^{\text{dl}} = I\left(\tilde{\mathbf{x}}_i^{\text{dl}}; \hat{\mathbf{x}}_i^{\text{dl}}\right) = g_i^{\text{dl}}\left(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}_i^{\text{dl}}\right) = \\ &\log_2 \det\left(\sum\nolimits_{k\in\mathcal{N}_U} \mathbf{E}_i^H \mathbf{Q}_{C,k}^{\text{dl}} \mathbf{E}_i + \mathbf{\Omega}_i^{\text{dl}}\right) - \log_2 \det\left(\mathbf{\Omega}_i^{\text{dl}}\right), \end{aligned} \tag{47}$$

where the elements of $\mathbf{E}_i \in \mathbb{C}^{n_E\times n_{E,i}}$ are filled with zeros except for the rows from $\sum_{j=1}^{i-1} n_{E,j} + 1$ to $\sum_{j=1}^{i} n_{E,j}$ being an identity matrix of size $n_{E,i} \times n_{E,i}$.

Similar to (41) for uplink, the downlink fronthaul latency τ_F^{dl} for given γ_i^{dl}, $i \in \mathcal{N}_E$, and τ_E^{dl} is computed as

$$\tau_F^{\text{dl}} = \max_{i\in\mathcal{N}_E} \frac{W^{\text{dl}}\tau_E^{\text{dl}}\gamma_i^{\text{dl}}}{C_F}. \tag{48}$$

Each EN i also encodes the edge computation output information for UE $k \in \mathcal{N}_{U,i}$ of $c_k b_{O,k}$ bits producing an encoded baseband signal $\mathbf{s}_{E,k}^{\text{dl}} \in \mathbb{C}^{n_{E,i}\times 1} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}_{E,k}^{\text{dl}})$. EN i then transmits a superposition of the locally encoded signals $\mathbf{s}_{E,k}^{\text{dl}}$, $k \in \mathcal{N}_{U,i}$, and the quantized signal $\hat{\mathbf{x}}_i^{\text{dl}}$, which was received on fronthaul, over the downlink channel to UEs. Thus, the signal $\mathbf{x}_i^{\text{dl}}$ transmitted by EN i is given as

$$\mathbf{x}_i^{\text{dl}} = \sum\nolimits_{k\in\mathcal{N}_{U,i}} \mathbf{s}_{E,k}^{\text{dl}} + \hat{\mathbf{x}}_i^{\text{dl}}. \tag{49}$$

With (49), the transmit power constraint (9) at EN i can be written as

$$\sum_{k\in\mathcal{N}_{U,i}} \text{tr}\left(\mathbf{Q}_{E,k}^{\text{dl}}\right) + \sum_{k\in\mathcal{N}_U} \text{tr}\left(\mathbf{E}_i^H \mathbf{Q}_{C,k}^{\text{dl}} \mathbf{E}_i\right) + \text{tr}\left(\mathbf{\Omega}_i^{\text{dl}}\right) \leq P^{\text{dl}}. \tag{50}$$

The first term in the left-hand side (LHS) measures the power of the signals $\{\mathbf{s}_{E,k}^{\text{dl}}\}_{k\in\mathcal{N}_{U,i}}$, which encode the computation output information processed by EN i. The sum of the second and third terms is the power of the signal $\hat{\mathbf{x}}_i^{\text{dl}}$, which is a quantized version of $\tilde{\mathbf{x}}_i^{\text{dl}}$ that encodes the signals $\{\mathbf{s}_{C,k}\}_{k\in\mathcal{N}_U}$ processed by the CP.

Each UE k detects the signals $\mathbf{s}_{E,k}^{\text{dl}}$ and $\mathbf{s}_{C,k}^{\text{dl}}$ based on the downlink received signal y_k^{dl}. The achievable rates of $\mathbf{s}_{E,k}^{\text{dl}}$ and $\mathbf{s}_{C,k}^{\text{dl}}$ are given as $R_{E,k}^{\text{dl}} = W^{\text{dl}} I(\mathbf{s}_{E,k}^{\text{dl}}; y_k^{\text{dl}})$ and $R_{C,k}^{\text{dl}} = W^{\text{dl}} I(\mathbf{s}_{C,k}^{\text{dl}}; y_k^{\text{dl}})$, respectively, with

$$I\left(\mathbf{s}_{E,k}^{\text{dl}}; y_k^{\text{dl}}\right) = f_{E,k}^{\text{dl}}\left(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}}\right) = \tag{51a}$$
$$\Psi\left(\mathbf{h}_{k,i_k}^{\text{dl}H}\mathbf{Q}_{E,k}^{\text{dl}}\mathbf{h}_{k,i_k}^{\text{dl}}, \begin{pmatrix} \sum_{l\in\mathcal{N}_U\setminus\{k\}} \mathbf{h}_{k,i_l}^{\text{dl}H}\mathbf{Q}_{E,l}^{\text{dl}}\mathbf{h}_{k,i_l}^{\text{dl}} \\ +\sum_{l\in\mathcal{N}_U} \mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,l}^{\text{dl}}\mathbf{h}_k^{\text{dl}} \\ +\sigma_{z,\text{dl}}^2 + \mathbf{h}_k^{\text{dl}H}\bar{\mathbf{\Omega}}^{\text{dl}}\mathbf{h}_k^{\text{dl}} \end{pmatrix}\right), \text{ and}$$
$$I\left(\mathbf{s}_{C,k}^{\text{dl}}; y_k^{\text{dl}}\right) = f_{C,k}^{\text{dl}}\left(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}}\right) = \tag{51b}$$
$$\Psi\left(\mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,k}^{\text{dl}}\mathbf{h}_k^{\text{dl}}, \begin{pmatrix} \sum_{l\in\mathcal{N}_U} \mathbf{h}_{k,i_l}^{\text{dl}H}\mathbf{Q}_{E,l}^{\text{dl}}\mathbf{h}_{k,i_l}^{\text{dl}} + \\ \sum_{l\in\mathcal{N}_U\setminus\{k\}} \mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,l}^{\text{dl}}\mathbf{h}_k^{\text{dl}} \\ +\sigma_{z,\text{dl}}^2 + \mathbf{h}_k^{\text{dl}H}\bar{\mathbf{\Omega}}^{\text{dl}}\mathbf{h}_k^{\text{dl}} \end{pmatrix}\right).$$

Here, we have defined $\mathbf{h}_k^{\text{dl}} = [\mathbf{h}_{k,1}^{\text{dl}H}\, \mathbf{h}_{k,2}^{\text{dl}H} \cdots \mathbf{h}_{k,N_E}^{\text{dl}H}]^H$ and $\bar{\mathbf{\Omega}}^{\text{dl}} = \text{diag}(\{\mathbf{\Omega}_i^{\text{dl}}\}_{i\in\mathcal{N}_E})$.

With the downlink rates described above, the latency τ_E^{dl} for downloading the output information on the downlink channel is given as

$$\tau_E^{\text{dl}} = \max_{k\in\mathcal{N}_U}\left\{\frac{c_k b_{O,k}}{W^{\text{dl}} f_{E,k}^{\text{dl}}(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}})}, \frac{(1-c_k) b_{O,k}}{W^{\text{dl}} f_{C,k}^{\text{dl}}(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}})}\right\}. \tag{52}$$

C. Total End-to-End Latency With C-RAN

The total end-to-end latency τ_T for completing the all the tasks within the described C-RAN architecture is modeled as

$$\tau_T = \tau_E^{\text{ul}} + \max\left\{\tau_E^{\text{exe}}, \tau_F^{\text{ul}} + \tau_C^{\text{exe}} + \tau_F^{\text{dl}}\right\} + \tau_E^{\text{dl}}, \tag{53}$$

where the fronthaul latency τ_F^{ul}, τ_F^{dl} and the edge latency τ_E^{ul}, τ_E^{dl} are defined in (41), (48), (44) and (52), respectively. Also, τ_E^{exe} and τ_C^{exe} represent the latency for executing the computation tasks at the ENs and CP which are are given as

$$\tau_E^{\text{exe}} = \max_{k\in\mathcal{N}_U} \tau_{E,i_k,k}^{\text{exe}} \quad \text{and} \quad \tau_C^{\text{exe}} = \max_{k\in\mathcal{N}_U} \tau_{C,k}^{\text{exe}}, \tag{54}$$

with $\tau_{E,i_k,k}^{\text{exe}}$ and $\tau_{C,k}^{\text{exe}}$ in (3) and (5).

D. Optimization

We aim at jointly optimizing the task splitting variables $\mathbf{c}$, the uplink $\{\mathbf{p}^{\text{ul}}, \mathbf{\Omega}^{\text{ul}}\}$ and downlink communication strategies $\{\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}}\}$ with the goal of minimizing the end-to-end latency τ_T in (53). The problem at hand can be stated as

$$\underset{\substack{\mathbf{p}\geq\mathbf{0}, \mathbf{c}\geq\mathbf{0}, \mathbf{Q}\succeq\mathbf{0}, \\ \mathbf{\Omega}\succeq\mathbf{0}, \mathbf{F}, \boldsymbol{\tau}, \mathbf{R}}}{\text{minimize}} \ \tau_E^{\text{ul}} + \max\left\{\tau_E^{\text{exe}}, \tau_F^{\text{ul}} + \tau_C^{\text{exe}} + \tau_F^{\text{dl}}\right\} + \tau_E^{\text{dl}} \tag{55a}$$
$$\text{s.t.} \quad \tau_E^{\text{ul}} \geq \frac{c_k b_{I,k}}{R_{E,k}^{\text{ul}}},\ k\in\mathcal{N}_U, \tag{55b}$$
$$\tau_E^{\text{ul}} \geq \frac{(1-c_k) b_{I,k}}{R_{C,k}^{\text{ul}}},\ k\in\mathcal{N}_U, \tag{55c}$$
$$\tau_F^{\text{ul}} \geq \frac{W^{\text{ul}}\tau_E^{\text{ul}}\, g_i^{\text{ul}}\left(\mathbf{p}^{\text{ul}}, \mathbf{\Omega}_i^{\text{ul}}\right)}{C_F},\ i\in\mathcal{N}_E, \tag{55d}$$
$$\tau_E^{\text{dl}} \geq \frac{c_k b_{O,k}}{R_{E,k}^{\text{dl}}},\ k\in\mathcal{N}_U, \tag{55e}$$
$$\tau_E^{\text{dl}} \geq \frac{(1-c_k) b_{O,k}}{R_{C,k}^{\text{dl}}},\ k\in\mathcal{N}_U, \tag{55f}$$
$$\tau_F^{\text{dl}} \geq \frac{W^{\text{dl}}\tau_E^{\text{dl}}\, g_i^{\text{dl}}\left(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}_i^{\text{dl}}\right)}{C_F},\ i\in\mathcal{N}_E, \tag{55g}$$
$$\text{(23f), (23g)}, \tag{55h}$$
$$R_{E,k}^{\text{ul}} \leq W^{\text{ul}} f_{E,k}^{\text{ul}}\left(\mathbf{p}^{\text{ul}}\right),\ k\in\mathcal{N}_U, \tag{55i}$$
$$R_{C,k}^{\text{ul}} \leq W^{\text{ul}} f_{C,k}^{\text{ul}}\left(\mathbf{p}^{\text{ul}}, \mathbf{\Omega}^{\text{ul}}\right),\ k\in\mathcal{N}_U, \tag{55j}$$
$$R_{E,k}^{\text{dl}} \leq W^{\text{dl}} f_{E,k}^{\text{dl}}\left(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}}\right),\ k\in\mathcal{N}_U \tag{55k}$$
$$R_{C,k}^{\text{dl}} \leq W^{\text{dl}} f_{C,k}^{\text{dl}}\left(\mathbf{Q}^{\text{dl}}, \mathbf{\Omega}^{\text{dl}}\right),\ k\in\mathcal{N}_U, \tag{55l}$$
$$\text{(23j)-(23m)}, \tag{55m}$$
$$p_{E,k}^{\text{ul}} + p_{C,k}^{\text{ul}} \leq P^{\text{ul}},\ k\in\mathcal{N}_U, \tag{55n}$$
$$\sum_{k\in\mathcal{N}_{U,i}} \text{tr}\left(\mathbf{Q}_{E,k}^{\text{dl}}\right) + \sum_{k\in\mathcal{N}_U} \text{tr}\left(\mathbf{E}_i^H\mathbf{Q}_{C,k}^{\text{dl}}\mathbf{E}_i\right) + \text{tr}\left(\mathbf{\Omega}_i^{\text{dl}}\right) \leq P^{\text{dl}},\ i\in\mathcal{N}_E, \tag{55o}$$
$$c_k \in [0,1],\ k\in\mathcal{N}_U. \tag{55p}$$

We note that it is more difficult to solve problem (55) than problems (23) and (31) for D-RAN, since (55) involves more optimization variables including the fronthaul quantization strategies $\mathbf{\Omega}^{\text{ul}}$ and $\mathbf{\Omega}^{\text{dl}}$; and the constraints (55d) and (55g) on the fronthaul latency have a more complicated form than (23c) and (23e) for D-RAN systems. To address these complications, we apply FP and matrix FP [27], [28] as in the methodology outlined above for D-RAN as well as the convex approximation method introduced in [19, Lem. 1].

To this end, we first replace the constraints (55h) with (24c) and (24d) which are convex for fixed $\lambda_{E,i_k,k}^{\text{exe}}$ and $\lambda_{C,k}^{\text{exe}}$ and become equivalent to (55h) when $\lambda_{E,i_k,k}^{\text{exe}}$ and $\lambda_{C,k}^{\text{exe}}$ are given as (25). Similarly, based on [27, Cor. 1], we consider the following constraints which are stricter than (55b), (55c), (55e) and (55f):

$$2\lambda_{E,k}^{\text{ul}}\sqrt{\tau_E^{\text{ul}}} - (\lambda_{E,k}^{\text{ul}})^2 c_k \geq \frac{b_{I,k}}{R_{E,k}^{\text{ul}}},\ k\in\mathcal{N}_U, \tag{56a}$$
$$2\lambda_{C,k}^{\text{ul}}\sqrt{\tau_E^{\text{ul}}} - (\lambda_{C,k}^{\text{ul}})^2 (1-c_k) \geq \frac{b_{I,k}}{R_{C,k}^{\text{ul}}}\ k\in\mathcal{N}_U, \tag{56b}$$
$$2\lambda_{E,k}^{\text{dl}}\sqrt{\tau_E^{\text{dl}}} - (\lambda_{E,k}^{\text{dl}})^2 c_k \geq \frac{b_{O,k}}{R_{E,k}^{\text{dl}}},\ k\in\mathcal{N}_U, \tag{56c}$$
$$2\lambda_{C,k}^{\text{dl}}\sqrt{\tau_E^{\text{dl}}} - (\lambda_{C,k}^{\text{dl}})^2 (1-c_k) \geq \frac{b_{O,k}}{R_{C,k}^{\text{dl}}},\ k\in\mathcal{N}_U. \tag{56d}$$

The above constraints become equivalent to (55b), (55c), (55e) and (55f) if

$$\lambda_{E,k}^{m} = \frac{\sqrt{\tau_E^m}}{c_k} \text{ and } \lambda_{C,k}^{m} = \frac{\sqrt{\tau_E^m}}{1-c_k}, \tag{57}$$

for $m \in \{\mathrm{ul}, \mathrm{dl}\}$.

Now, we discuss the non-convex constraints (55d) and (55g). Using the epigraph form, the constraint (55d) can be restated as

$$\tau_F^{\mathrm{ul}} \geq \frac{W^{\mathrm{ul}} \tau_E^{\mathrm{ul}} \gamma_i^{\mathrm{ul}}}{C_F}, \ i \in \mathcal{N}_E, \text{ and} \tag{58a}$$

$$\gamma_i^{\mathrm{ul}} \geq g_i^{\mathrm{ul}}\left(\mathbf{p}^{\mathrm{ul}}, \mathbf{\Omega}^{\mathrm{ul}}\right), \ i \in \mathcal{N}_E. \tag{58b}$$

From [27, Cor. 1] and [19, Lem. 1], the following constraints are stricter than (58):

$$\frac{W^{\mathrm{ul}}\gamma_i^{\mathrm{ul}}}{C_F} \leq 2\alpha^{\mathrm{ul}}\sqrt{\tau_F^{\mathrm{ul}}} - (\alpha^{\mathrm{ul}})^2 \tau_E^{\mathrm{ul}}, \ i \in \mathcal{N}_E, \text{ and} \tag{59a}$$

$$\gamma_i^{\mathrm{ul}} \geq \log_2 \det\left(\mathbf{\Sigma}_i^{\mathrm{ul}}\right) + \frac{1}{\ln 2} \times \mathrm{tr}\left(\mathbf{\Sigma}_i^{\mathrm{ul}-1}\left(\begin{array}{c}\sum_{k\in\mathcal{N}_U\setminus\mathcal{N}_{U,i}} p_{E,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}H} \\ + \sum_{k\in\mathcal{N}_U} p_{C,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}H} \\ + \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \mathbf{\Omega}_i^{\mathrm{ul}}\end{array}\right)\right) - \frac{n_{E,i}}{\ln 2} - \log_2 \det\left(\mathbf{\Omega}_i^{\mathrm{ul}}\right), \ i \in \mathcal{N}_E. \tag{59b}$$

If we fix the auxiliary variables α^{ul} and $\mathbf{\Sigma}_i^{\mathrm{ul}}$, the constraints (59) are convex. Also, they become equivalent to (58) if the auxiliary variables α^{ul} and $\mathbf{\Sigma}_i^{\mathrm{ul}}$ are given as

$$\alpha^{\mathrm{ul}} = \frac{\sqrt{\tau_F^{\mathrm{ul}}}}{\tau_E^{\mathrm{ul}}}, \text{ and} \tag{60a}$$

$$\mathbf{\Sigma}_i^{\mathrm{ul}} = \sum_{k\in\mathcal{N}_U\setminus\mathcal{N}_{U,i}} p_{E,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}H} + \sum_{k\in\mathcal{N}_U} p_{C,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}} \mathbf{h}_{i,k}^{\mathrm{ul}H} + \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \mathbf{\Omega}_i^{\mathrm{ul}}. \tag{60b}$$

Similarly, instead of (55g) for downlink, we consider the following stricter constraints:

$$\frac{W^{\mathrm{dl}}\gamma_i^{\mathrm{dl}}}{C_F} \leq 2\alpha^{\mathrm{dl}}\sqrt{\tau_F^{\mathrm{dl}}} - (\alpha^{\mathrm{dl}})^2 \tau_E^{\mathrm{dl}}, \ i \in \mathcal{N}_E, \text{ and} \tag{61a}$$

$$\gamma_i^{\mathrm{dl}} \geq \log_2 \det\left(\mathbf{\Sigma}_i^{\mathrm{dl}}\right) + \frac{1}{\ln 2} \times \mathrm{tr}\left(\mathbf{\Sigma}_i^{\mathrm{dl}-1}\left(\sum_{k\in\mathcal{N}_U} \mathbf{E}_i^H \tilde{\mathbf{Q}}_{C,k}^{\mathrm{dl}} \tilde{\mathbf{Q}}_{C,k}^{\mathrm{dl}H} \mathbf{E}_i + \mathbf{\Omega}_i^{\mathrm{dl}}\right)\right) - \frac{n_{E,i}}{\ln 2} - \log_2 \det\left(\mathbf{\Omega}_i^{\mathrm{dl}}\right), \ i \in \mathcal{N}_E. \tag{61b}$$

The above constraints are equivalent to (55g) if

$$\alpha^{\mathrm{dl}} = \frac{\sqrt{\tau_F^{\mathrm{dl}}}}{\tau_E^{\mathrm{dl}}}, \text{ and} \tag{62a}$$

$$\mathbf{\Sigma}_i^{\mathrm{dl}} = \sum_{k\in\mathcal{N}_U} \mathbf{E}_i^H \tilde{\mathbf{Q}}_{C,k}^{\mathrm{dl}} \tilde{\mathbf{Q}}_{C,k}^{\mathrm{dl}H} \mathbf{E}_i + \mathbf{\Omega}_i^{\mathrm{dl}}. \tag{62b}$$

Algorithm 3 Alternating optimization algorithm that tackles problem (55)

1. Initialize $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\tau}, \mathbf{R}\}$ as arbitrary matrices/values that satisfy the constraints (55b)-(55l), and set $t \leftarrow 1$.
2. Calculate the total latency τ_T in (53) with the initialized $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\tau}, \mathbf{R}\}$, and set $\tau_T^{(0)} \leftarrow \tau_T$.
3. Set $\{\boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta}\}$ according to (25), (57), (60), (62) and (64).
4. Update $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{R}\}$ as a solution of the convex problem which is obtained from (55) by replacing the constraints (55b)-(55l) with (24c), (24d), (56), (59), (61) and (63), and fixing the variables $\{\boldsymbol{\lambda}, \boldsymbol{\alpha}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta}\}$.
5. Calculate the total latency τ_T in (53) with the updated $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \mathbf{R}\}$, and set $\tau_T^{(t)} \leftarrow \tau_T$.
6. Stop if $|\tau_T^{(t)} - \tau_T^{(t-1)}| \leq \delta$ or $t > t_{\max}$. Otherwise, set $t \leftarrow t+1$ and go back to Step 3.

Lastly, using [28, Cor. 1], we replace the remaining non-convex constraints (55i)-(55l) with the following stricter constraints:

$$\frac{R_{E,k}^{\mathrm{ul}}}{W^{\mathrm{ul}}} \leq \phi\left(\begin{array}{c}\mathbf{\Gamma}_{E,k}^{\mathrm{ul}}, \mathbf{\Theta}_{E,k}^{\mathrm{ul}}, \tilde{p}_{E,k}^{\mathrm{ul}} \mathbf{h}_{i_k,k}^{\mathrm{ul}}, \\ \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \sum_{l\in\mathcal{N}_U\setminus\{k\}} p_{E,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}H} \\ + \sum_{l\in\mathcal{N}_U} p_{C,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}} \mathbf{h}_{i_k,l}^{\mathrm{ul}H}\end{array}\right), \tag{63a}$$

$$\frac{R_{C,k}^{\mathrm{ul}}}{W^{\mathrm{ul}}} \leq \phi\left(\begin{array}{c}\mathbf{\Gamma}_{C,k}^{\mathrm{ul}}, \mathbf{\Theta}_{C,k}^{\mathrm{ul}}, \tilde{p}_{C,k}^{\mathrm{ul}} \mathbf{h}_k^{\mathrm{ul}}, \\ \sigma_{z,\mathrm{ul}}^2 \mathbf{I} + \bar{\mathbf{\Omega}}^{\mathrm{ul}} + \sum_{l\in\mathcal{N}_U} p_{E,l}^{\mathrm{ul}} \tilde{\mathbf{h}}_l^{\mathrm{ul}} \tilde{\mathbf{h}}_l^{\mathrm{ul}H} \\ + \sum_{l\in\mathcal{N}_U\setminus\{k\}} p_{C,l}^{\mathrm{ul}} \mathbf{h}_l^{\mathrm{ul}} \mathbf{h}_l^{\mathrm{ul}H}\end{array}\right), \tag{63b}$$

$$\frac{R_{E,k}^{\mathrm{dl}}}{W^{\mathrm{dl}}} \leq \phi\left(\begin{array}{c}\mathbf{\Gamma}_{E,k}^{\mathrm{dl}}, \mathbf{\Theta}_{E,k}^{\mathrm{dl}}, \mathbf{h}_{k,i_k}^{\mathrm{dl}H} \tilde{\mathbf{Q}}_{E,k}^{\mathrm{dl}}, \\ \sigma_{z,\mathrm{dl}}^2 + \mathbf{h}_k^{\mathrm{dl}H} \bar{\mathbf{\Omega}}^{\mathrm{dl}} \mathbf{h}_k^{\mathrm{dl}} + \\ \sum_{l\in\mathcal{N}_U\setminus\{k\}} \mathbf{h}_{k,i_l}^{\mathrm{dl}H} \mathbf{Q}_{E,l} \mathbf{h}_{k,i_l}^{\mathrm{dl}} \\ + \sum_{l\in\mathcal{N}_U} \mathbf{h}_k^{\mathrm{dl}H} \mathbf{Q}_{C,l} \mathbf{h}_k^{\mathrm{dl}}\end{array}\right), \text{ and} \tag{63c}$$

$$\frac{R_{C,k}^{\mathrm{dl}}}{W^{\mathrm{dl}}} \leq \phi\left(\begin{array}{c}\mathbf{\Gamma}_{C,k}^{\mathrm{dl}}, \mathbf{\Theta}_{C,k}^{\mathrm{dl}}, \mathbf{h}_k^{\mathrm{dl}H} \tilde{\mathbf{Q}}_{C,k}^{\mathrm{dl}}, \\ \sigma_{z,\mathrm{dl}}^2 + \mathbf{h}_k^{\mathrm{dl}H} \bar{\mathbf{\Omega}}^{\mathrm{dl}} \mathbf{h}_k^{\mathrm{dl}} + \\ \sum_{l\in\mathcal{N}_U} \mathbf{h}_{k,i_l}^{\mathrm{dl}H} \mathbf{Q}_{E,l} \mathbf{h}_{k,i_l}^{\mathrm{dl}} \\ + \sum_{l\in\mathcal{N}_U\setminus\{k\}} \mathbf{h}_k^{\mathrm{dl}H} \mathbf{Q}_{C,l} \mathbf{h}_k^{\mathrm{dl}}\end{array}\right), \tag{63d}$$

for $k \in \mathcal{N}_U$. The above constraints are equivalent to (55i)-(55l) if the variables $\mathbf{\Gamma} \triangleq \{\Gamma_{E,k}^{\mathrm{ul}}, \Gamma_{C,k}^{\mathrm{ul}}, \mathbf{\Gamma}_{E,k}^{\mathrm{dl}}, \mathbf{\Gamma}_{C,k}^{\mathrm{dl}}\}_{k\in\mathcal{N}_U}$ and $\mathbf{\Theta} \triangleq \{\mathbf{\Theta}_{E,k}^{\mathrm{ul}}, \mathbf{\Theta}_{C,k}^{\mathrm{ul}}, \mathbf{\Theta}_{E,k}^{\mathrm{dl}}, \mathbf{\Theta}_{C,k}^{\mathrm{dl}}\}_{k\in\mathcal{N}_U}$ are given as (64) at the bottom of p. 11.

Based on the discussed inequalities (24c), (24d), (56), (59), (61), and (63) that restate the non-convex constraints (55b)-(55l) of problem (55), we propose an iterative algorithm that alternately optimizes $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\tau}, \mathbf{R}\}$ and $\{\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta}\}$. When optimizing $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\tau}, \mathbf{R}\}$, we tackle the convex problem which is obtained from (55) by replacing the constraints (55b)-(55l) with (24c), (24d), (56), (59), (61) and (63), and fixing the variables $\{\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta}\}$. For fixed $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}, \boldsymbol{\tau}, \mathbf{R}\}$, the optimal variables $\{\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{\Sigma}, \mathbf{\Gamma}, \mathbf{\Theta}\}$ are obtained as (25), (57), (60), (62) and (64). The detailed algorithm is described in Algorithm 3. The solution obtained by Algorithm 3 is a locally optimal solution due to the non-convexity of the problem (55). In Sec. V, we initialize $\{\mathbf{p}, \mathbf{c}\}$ as $p_{E,k}^{\mathrm{ul}} \leftarrow P^{\mathrm{ul}}$, $p_{C,k}^{\mathrm{ul}} \leftarrow P^{\mathrm{ul}}$ and $c_k \leftarrow 1/2$ for $k \in \mathcal{N}_U$. To initialize the covariance matrices

of downlink signals $\mathbf{Q}$ and quantization noise signals $\mathbf{\Omega}$, we first set

$$\mathbf{Q}_{E,k} \leftarrow \mathbf{V}_{E,k}\mathbf{V}_{E,k}^H,\ k \in \mathcal{N}_{U,i}, i \in \mathcal{N}_E, \tag{65a}$$

$$\mathbf{Q}_{C,k} \leftarrow \mathbf{V}_{C,k}\mathbf{V}_{C,k}^H,\ k \in \mathcal{N}_U, \tag{65b}$$

$$\mathbf{\Omega}_i \leftarrow \mathbf{V}_{\Omega,i}\mathbf{V}_{\Omega,i}^H,\ i \in \mathcal{N}_E, \tag{65c}$$

where the elements of $\mathbf{V}_{E,k} \in \mathbb{C}^{n_{E,i}\times n_{E,i}}$, $\mathbf{V}_{C,k} \in \mathbb{C}^{n_E \times n_E}$ and $\mathbf{V}_{\Omega,k} \in \mathbb{C}^{n_{E,i}\times n_{E,i}}$ follow $\mathcal{CN}(0,1)$. The covariance matrices obtained in (65) may not satisfy the power constraints (50). To resolve this issue, we repeatedly multiply a scalar $\eta < 1$ to the matrices $\mathbf{Q}$ and $\mathbf{\Omega}$ until the constraints (50) are satisfied. In the simulation, we set $\eta = 1/2$. Once the variables $\{\mathbf{p}, \mathbf{c}, \mathbf{Q}, \mathbf{\Omega}\}$ are fixed, the rate variables $\mathbf{R}$ can be computed using (37), (43) and (51), and the latency variables $\boldsymbol{\tau}$ are initialized as (41), (44), (48), and (52).

As discussed in Sec. III, the complexity of Algorithm 3 is given by the number of iterations multiplied by the complexity of solving the convex problem at Step 4. The complexity of the latter is upper bounded by $\mathcal{O}(n(n^3+M)\log(1/\epsilon))$ [32, p. 4], where the numbers n and M equal $n = N_U(4\tilde{n}_E^2(N_E^2+1)+10)+N_E(8\tilde{n}_E^2+2)+6$ and $M = \left(8\tilde{n}_E^2 N_U + D_{\tilde{n}_E}\right)N_E + 4N_U N_E^2\tilde{n}_E^2(8N_E\tilde{n}_E+3N_U)+50N_U+5N_E\tilde{n}_E$, respectively. Here D_L is defined as the number of arithmetic operations needed to calculate the determinant of an $L\times L$ matrix, which is given as $D_L = \mathcal{O}(L^3)$ with Gaussian elimination [37, p. 1]. We discuss the convergence rate of Algorithm 3 in Sec. V.

V. Numerical Results

In this section, we validate via numerical results the performance gain of the proposed C-RAN architecture as compared to the D-RAN reference system. We assume that the locations of N_U UEs and N_E ENs are independently and uniformly sampled from a square area with side length of 500 m. We impose the minimum separation of 10 m between any pair of UE and EN. We consider a path-loss model $\rho_0(d/d_0)^{-\eta}$ [38], [39], where ρ_0 is the path-loss at a reference distance d_0, d denotes the distance between the transmitting and receiving nodes, and η is the path-loss exponent. We set $d_0 = 30$ m, $\rho_0 = 10$ dB and $\eta = 3$, and assume independent Rayleigh small-scale fading channel model for all the channel coefficients. We consider a symmetric system between uplink and downlink with $\text{SNR}_{\max}^{\text{ul}} = \text{SNR}_{\max}^{\text{dl}} = \text{SNR}_{\max}$, $W^{\text{ul}} = W^{\text{dl}} = W$, and $C_F^{\text{ul}} = C_F^{\text{dl}} = C_F$. The computation capabilities of CP and ENs are set to $F_C = 10^{11}$ [4] and $F_{E,i} \in \{1.0, 2.5\}\times 10^{10}$ [13], [40], respectively, unless stated otherwise. We also assume that there are $b_{I,k} = b_{O,k} = 10^6$ input and output bits for each UE and that the task of each UE k requires $V_k = 700$ CPU cycles per input bit [8]. To solve the convex problems at Step 4 of Algorithms 1, 2 and 3, CVX software [31] with SDPT3 solver [41] is adopted. Without claim of optimality, we associate each UE k with the closest EN, so that i_k is set to

$$i_k = \arg\min_{i\in\mathcal{N}_E} \text{dist}_{i,k}, \tag{66}$$

with $\text{dist}_{i,k}$ represents the geographical distance between UE k and EN i.

$$\Gamma_{E,k}^{\text{ul}} = p_{E,k}^{\text{ul}}\mathbf{h}_{i_k,k}^{\text{ul}H}\left(\sigma_{z,\text{ul}}^2\mathbf{I} + \sum_{l\in\mathcal{N}_U\setminus\{k\}} p_{E,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}H} + \sum_{l\in\mathcal{N}_U} p_{C,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}H}\right)^{-1}\mathbf{h}_{i_k,k}^{\text{ul}}, \tag{64a}$$

$$\mathbf{\Theta}_{E,k}^{\text{ul}} = \tilde{p}_{E,k}^{\text{ul}}\left(\sigma_{z,\text{ul}}^2\mathbf{I} + \sum_{l\in\mathcal{N}_U} p_{E,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}H} + \sum_{l\in\mathcal{N}_U} p_{C,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}}\mathbf{h}_{i_k,l}^{\text{ul}H}\right)^{-1}\mathbf{h}_{i_k,k}^{\text{ul}}, \tag{64b}$$

$$\Gamma_{C,k}^{\text{ul}} = p_{C,k}^{\text{ul}}\mathbf{h}_k^{\text{ul}H}\left(\sigma_{z,\text{ul}}^2\mathbf{I} + \bar{\mathbf{\Omega}}^{\text{ul}} + \sum_{l\in\mathcal{N}_U} p_{E,l}^{\text{ul}}\tilde{\mathbf{h}}_l^{\text{ul}}\tilde{\mathbf{h}}_l^{\text{ul}H} + \sum_{l\in\mathcal{N}_U\setminus\{k\}} p_{C,l}^{\text{ul}}\mathbf{h}_l^{\text{ul}}\mathbf{h}_l^{\text{ul}H}\right)^{-1}\mathbf{h}_k^{\text{ul}}, \tag{64c}$$

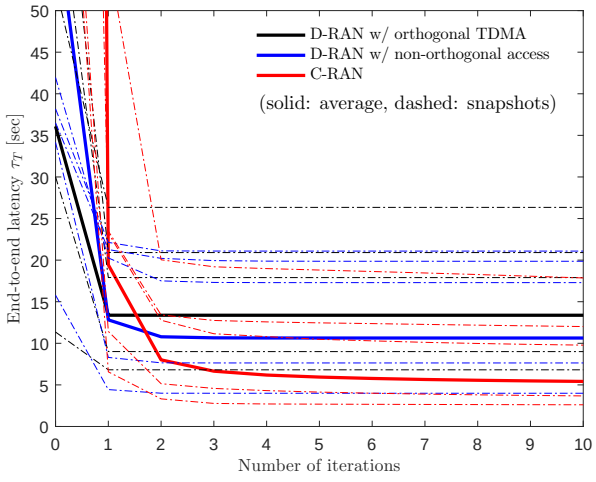
$$\mathbf{\Theta}_{C,k}^{\text{ul}} = \tilde{p}_{C,k}^{\text{ul}}\left(\sigma_{z,\text{ul}}^2\mathbf{I} + \bar{\mathbf{\Omega}}^{\text{ul}} + \sum_{l\in\mathcal{N}_U} p_{E,l}^{\text{ul}}\tilde{\mathbf{h}}_l^{\text{ul}}\tilde{\mathbf{h}}_l^{\text{ul}H} + \sum_{l\in\mathcal{N}_U} p_{C,l}^{\text{ul}}\mathbf{h}_l^{\text{ul}}\mathbf{h}_l^{\text{ul}H}\right)^{-1}\mathbf{h}_k^{\text{ul}}, \tag{64d}$$

$$\mathbf{\Gamma}_{E,k}^{\text{dl}} = \tilde{\mathbf{Q}}_{E,k}^{\text{dl}H}\mathbf{h}_{k,i_k}^{\text{dl}}\left(\sigma_{z,\text{dl}}^2 + \mathbf{h}_k^{\text{dl}H}\bar{\mathbf{\Omega}}^{\text{dl}}\mathbf{h}_k^{\text{dl}} + \sum_{l\in\mathcal{N}_U\setminus\{k\}}\mathbf{h}_{k,i_l}^{\text{dl}H}\mathbf{Q}_{E,l}\mathbf{h}_{k,i_l}^{\text{dl}} + \sum_{l\in\mathcal{N}_U}\mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,l}\mathbf{h}_k^{\text{dl}}\right)^{-1}\mathbf{h}_{k,i_k}^{\text{dl}H}\tilde{\mathbf{Q}}_{E,k}^{\text{dl}}, \tag{64e}$$

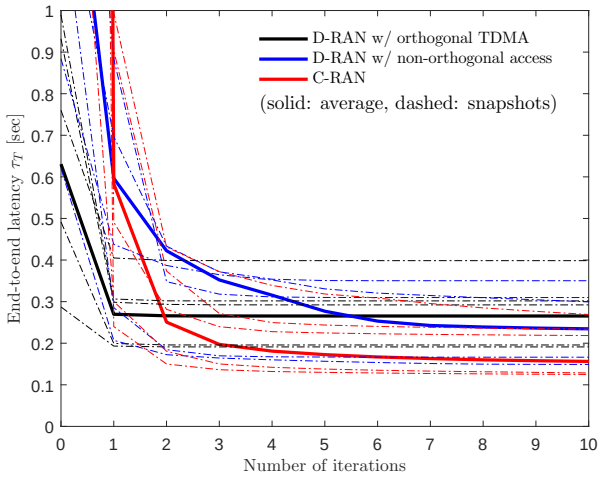
$$\mathbf{\Theta}_{E,k}^{\text{dl}} = \left(\sigma_{z,\text{dl}}^2 + \mathbf{h}_k^{\text{dl}H}\bar{\mathbf{\Omega}}^{\text{dl}}\mathbf{h}_k^{\text{dl}} + \sum_{l\in\mathcal{N}_U}\mathbf{h}_{k,i_l}^{\text{dl}H}\mathbf{Q}_{E,l}\mathbf{h}_{k,i_l}^{\text{dl}} + \sum_{l\in\mathcal{N}_U}\mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,l}\mathbf{h}_k^{\text{dl}}\right)^{-1}\mathbf{h}_{k,i_k}^{\text{dl}H}\tilde{\mathbf{Q}}_{E,k}^{\text{dl}}, \tag{64f}$$

$$\mathbf{\Gamma}_{C,k}^{\text{dl}} = \tilde{\mathbf{Q}}_{C,k}^{\text{dl}H}\mathbf{h}_k^{\text{dl}}\left(\sigma_{z,\text{dl}}^2 + \mathbf{h}_k^{\text{dl}H}\bar{\mathbf{\Omega}}^{\text{dl}}\mathbf{h}_k^{\text{dl}} + \sum_{l\in\mathcal{N}_U}\mathbf{h}_{k,i_l}^{\text{dl}H}\mathbf{Q}_{E,l}\mathbf{h}_{k,i_l}^{\text{dl}} + \sum_{l\in\mathcal{N}_U\setminus\{k\}}\mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,l}\mathbf{h}_k^{\text{dl}}\right)^{-1}\mathbf{h}_k^{\text{dl}H}\tilde{\mathbf{Q}}_{C,k}^{\text{dl}}, \tag{64g}$$

$$\mathbf{\Theta}_{C,k}^{\text{dl}} = \left(\sigma_{z,\text{dl}}^2 + \mathbf{h}_k^{\text{dl}H}\bar{\mathbf{\Omega}}^{\text{dl}}\mathbf{h}_k^{\text{dl}} + \sum_{l\in\mathcal{N}_U}\mathbf{h}_{k,i_l}^{\text{dl}H}\mathbf{Q}_{E,l}\mathbf{h}_{k,i_l}^{\text{dl}} + \sum_{l\in\mathcal{N}_U}\mathbf{h}_k^{\text{dl}H}\mathbf{Q}_{C,l}\mathbf{h}_k^{\text{dl}}\right)^{-1}\mathbf{h}_k^{\text{dl}H}\tilde{\mathbf{Q}}_{C,k}^{\text{dl}}. \tag{64h}$$



(a) $\text{SNR}_{\max} = 0$ dB



(b) $\text{SNR}_{\max} = 20$ dB

Figure 2. End-to-end latency τ_T versus the number of iterations ($N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 20$ MHz, $C_F = 1$ Gbps, $F_{E,i} = 10^{10}$ and $\text{SNR}_{\max} \in \{0, 20\}$ dB).

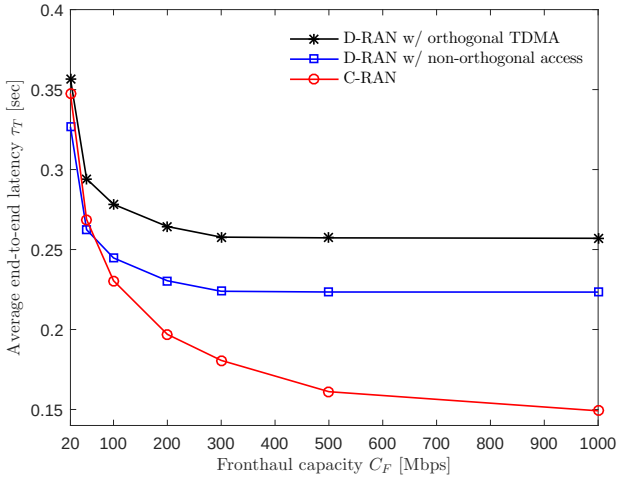


Figure 3. Average end-to-end latency τ_T versus the fronthaul capacity C_F ($N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 20$ MHz, $F_{E,i} = 10^{10}$ and $\text{SNR}_{\max} = 20$ dB).

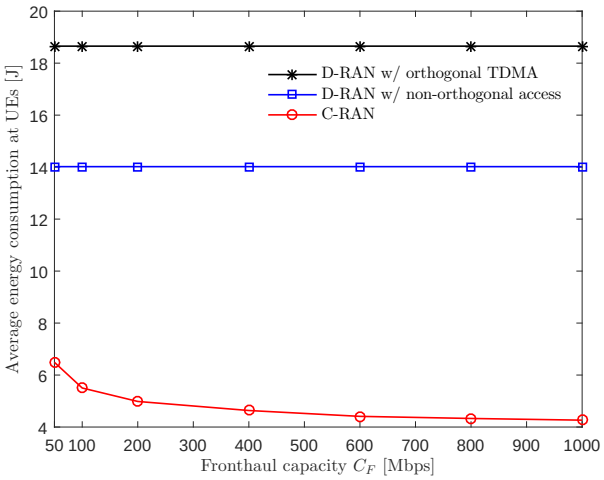


Figure 4. Average energy consumption at UEs versus the fronthaul capacity C_F ($N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 20$ MHz, $F_{E,i} = 10^{10}$ and $\text{SNR}_{\max} = 20$ dB).

A. Convergence of the Proposed Algorithm

The convergence rate of FP is analyzed in [27] with a focus on single-ratio problems, and reference [28] discusses the convergence rate of matrix FP via numerical examples. Similar to [28], we provide numerical evidence of the fast convergence of the proposed algorithms in Fig. 2. In the figure, we plot the end-to-end latency τ_T of D-RAN and C-RAN schemes versus the number of iterations for $N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 20$ MHz, $C_F = 1$ Gbps, $F_{E,i} = 10^{10}$ and $\text{SNR}_{\max} \in \{0, 20\}$ dB. We plot both the snapshots and average latency, where the latter is averaged over 100 channel samples. The figure shows that, regardless of the SNR, the proposed algorithms converge reliably within a few iterations. We leave the analysis of the convergence rate of the proposed algorithms for future work. Throughout the following experiments, we set the threshold value for convergence as $\delta = 10^{-4}$ and limit the maximum number of iterations to $t_{\max} = 30$.

B. Performance Gains of the C-RAN Architecture

In this subsection, we investigate the performance gains of the C-RAN architecture introduced in Sec. IV for collaborative cloud and edge mobile computing as compared to benchmark D-RAN systems described in Sec. III. To this end, in Fig. 3, we plot the average end-to-end latency τ_T versus the fronthaul capacity C_F for $N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 20$ MHz, $F_{E,i} = 10^{10}$ and $\text{SNR}_{\max} = 20$ dB. The figure shows that deploying C-RAN architecture is not advantageous when the fronthaul capacity C_F is small due to the large latency caused by the fronthaul transmission. However, as C_F increases, the C-RAN scheme significantly outperforms the benchmark D-RAN schemes, since it enables more effective interference management by means of centralized encoding and decoding at CP.

In Fig. 4, we examine the energy consumption at UEs under

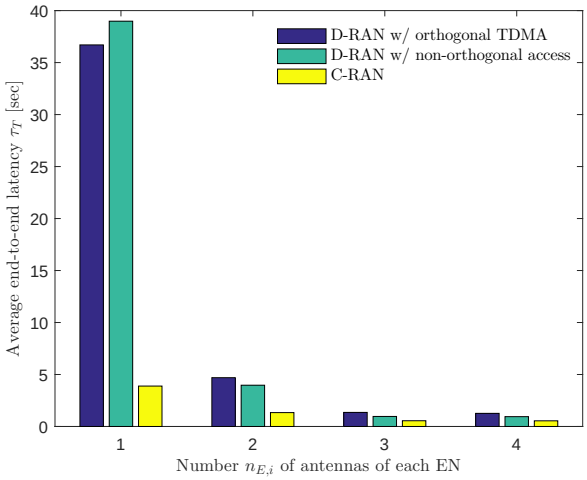


Figure 5. Average end-to-end latency τ_T versus the number $n_{E,i}$ of antennas of each EN ($N_U = 3$, $N_E = 2$, $W = 20$ MHz, $F_{E,i} = 10^{10}$, $C_F = 3$ Gbps and $\text{SNR}_{\max} = 5$ dB).

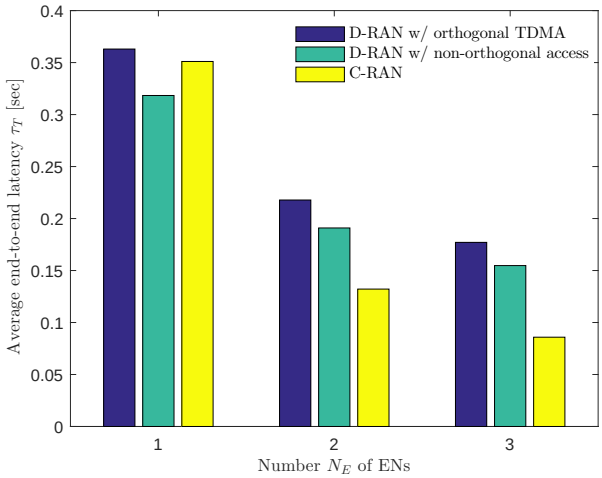


Figure 6. Average end-to-end latency τ_T versus the number N_E of ENs ($N_U = 8$, $n_{E,i} = 2$, $W = 50$ MHz, $F_{E,i} = 2.5 \times 10^{10}$, $C_F = 2$ Gbps and $\text{SNR}_{\max} = 20$ dB).

the same set-up considered in Fig. 3. We calculate the energy consumption at UE k as $E_k = E_k^{\text{ul}} + E_k^{\text{dl}}$, where the uplink and downlink energy expenditures are defined as $E_k^{\text{ul}} = \tau_{E,k}^{\text{ul}} \tilde{p}_k^{\text{ul}}$ and $E_k^{\text{dl}} = \tau_{E,k}^{\text{dl}} d_k^{\text{dl}}$, respectively. Here, d_k^{dl} indicates the mobile receiving energy expenditure per second in downlink, and is set to $d_k = 0.625$ J/s as in [13]. The uplink transmit power $\tilde{p}_k^{\text{ul}}$ of UE k is respectively given as $\tilde{p}_k^{\text{ul}} = p_k^{\text{ul}}$ and $\tilde{p}_k^{\text{ul}} = p_{E,k}^{\text{ul}} + p_{C,k}^{\text{ul}}$ for the D-RAN and C-RAN systems. Unlike D-RAN, the energy consumption of UEs with C-RAN decreases with C_F. This is because the ENs and CP can exchange quantized baseband signals of better resolution for larger C_F, and hence the latency on edge links becomes lower.

Fig. 5 plots the average end-to-end latency τ_T with respect to the number $n_{E,i}$ of antennas of each EN for $N_U = 3$, $N_E = 2$, $W = 20$ MHz, $F_{E,i} = 10^{10}$, $C_F = 3$ Gbps and $\text{SNR}_{\max} = 5$ dB. Comparing the performance of D-RAN with different access techniques, we see that TDMA shows a lower latency than non-orthogonal access when the ENs use a small number of antennas. However, when the ENs are equipped with sufficiently many antennas, the non-orthogonal scheme outperforms the TDMA scheme, since the co-channel interference signals can be suppressed by local array processing at the ENs. In this case, each EN can suppress interference signals only with local processing, and hence C-RAN does not provide performance benefits, while significant gains are observed for lower values of $n_{E,i}$.

In Fig. 6, we plot the average end-to-end latency τ_T versus the number N_E of ENs for $N_U = 8$, $n_{E,i} = 2$, $W = 50$ MHz, $F_{E,i} = 2.5 \times 10^{10}$, $C_F = 2$ Gbps and $\text{SNR}_{\max} = 20$ dB. When the network has a single EN, i.e., $N_E = 1$, there is no advantage of deploying the C-RAN architecture in Sec. IV compared to D-RAN in Sec. III. This is because the noise signals caused by fronthaul quantization degrade the spectral efficiency for both uplink and downlink. However, as N_E increases, C-RAN shows significantly improved latency performance than the D-RAN schemes. These gains are achieved by the centralized signal processing at the CP on behalf of the connected ENs, which enables effective interference management.

C. Performance Gains of Collaborative Cloud-Edge Computing

In this subsection, we study the performance gains of the collaborative cloud and edge computing system with optimized computational resource allocation as compared to benchmark schemes that rely only on edge computing (i.e., by setting $c_k = 1$ for all $k \in \mathcal{N}_U$) or cloud computing (i.e., $c_k = 0$ for all $k \in \mathcal{N}_U$). Note that the optimization of these benchmark schemes can be addressed by adopting the proposed algorithm with minor modifications. For reference, we also evaluate the performance of a hybrid strategy that selects between the two benchmark schemes. We adopt the optimized C-RAN architecture in Sec. IV for all cases except for edge computing, for which the C-RAN system is not applicable and hence we select D-RAN with non-orthogonal multiple access.

In Fig. 7, we plot the average end-to-end latency τ_T versus the fronthaul capacity C_F for $N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 50$ MHz, $F_{E,i} = 2.5 \times 10^{10}$ and $\text{SNR}_{\max} = 10$ dB. Since edge computing does not utilize the fronthaul links, its performance is not affected by C_F. In contrast, the latency of cloud computing scheme decreases as C_F increases. While selecting between edge and cloud computing schemes does not yield significant benefits, the proposed collaborative cloud and edge scheme achieves notable gains, particularly in the intermediate regime of C_F.

In Fig. 8, we plot the average end-to-end latency τ_T versus the maximum SNR for $N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 100$ MHz, $F_{E,i} = 2.5 \times 10^{10}$ and $C_F = 250$ Mbps. The figure shows that, although increased SNR levels are beneficial for all the schemes, the performance of cloud computing is more significantly affected by the SNR than that of edge computing. This is because the edge latency of edge computing is limited by interference, and hence its performance saturates as the

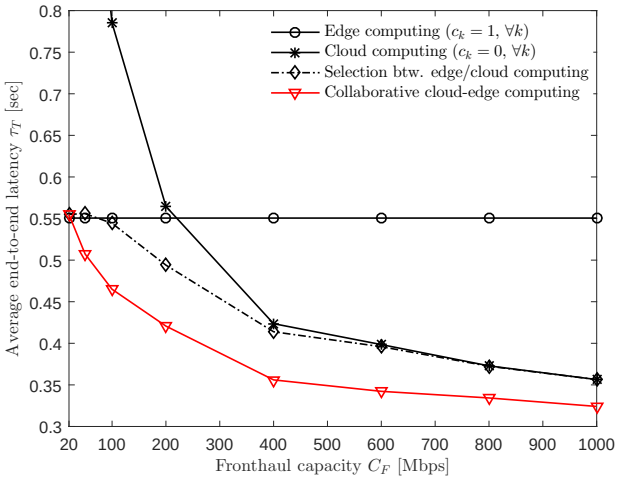


Figure 7. Average end-to-end latency τ_T versus the fronthaul capacity C_F ($N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 50$ MHz, $F_{E,i} = 2.5 \times 10^{10}$ and $\text{SNR}_{\max} = 10$ dB).

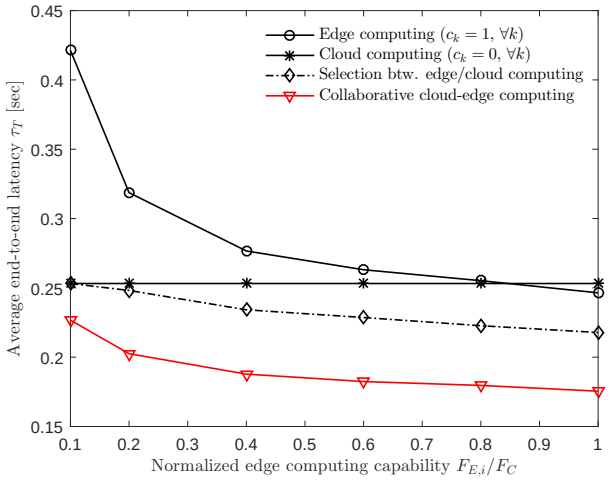


Figure 9. Average end-to-end latency τ_T versus the normalized edge computing capability $F_{E,i}/F_C$ ($N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 100$ MHz, $C_F = 500$ Mbps, $\text{SNR}_{\max} = 10$ dB and $F_C = 10^{11}$).

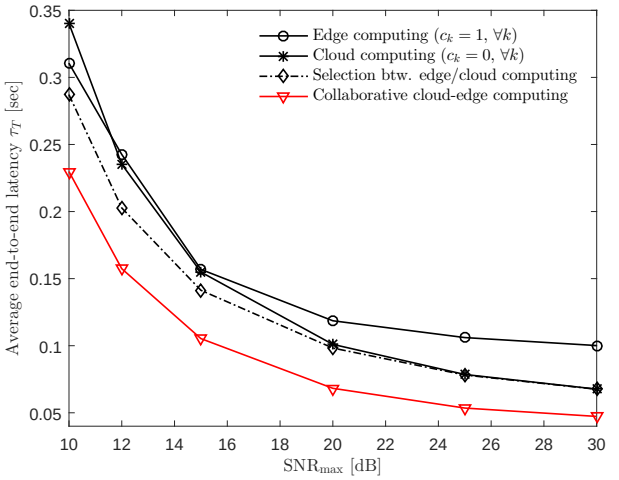


Figure 8. Average end-to-end latency τ_T versus the maximum SNR ($N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 100$ MHz, $F_{E,i} = 2.5 \times 10^{10}$ and $C_F = 250$ Mbps).

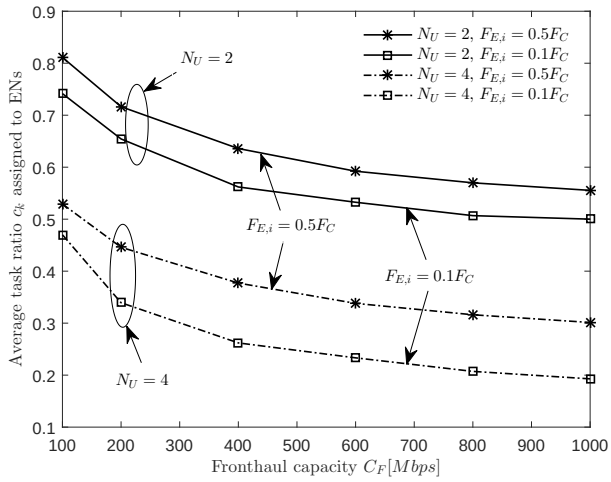


Figure 10. Average task ratio c_k assigned to ENs versus the fronthaul capacity C_F ($N_U \in \{2, 4\}$, $N_E = 2$, $n_{E,i} = 1$, $W = 100$ MHz and $F_{E,i} \in \{0.1, 0.5\} \times 10^{10}$).

SNR increases. The performance of the C-RAN scheme is instead limited by the fronthaul capacity as SNR grows larger.

Fig. 9 plots the average end-to-end latency τ_T by varying the edge computing capability $F_{E,i}$ normalized by F_C for $N_U = 4$, $N_E = 2$, $n_{E,i} = 2$, $W = 100$ MHz, $C_F = 500$ Mbps, $\text{SNR}_{\max} = 10$ dB and $F_C = 10^{11}$. When $F_{E,i}$ is too small, it is desired to choose $c_k = 0$ for all $k \in \mathcal{N}_U$ so that all the tasks are offloaded to the CP. As $F_{E,i}$ increases, offloading some tasks to ENs can improve the performance, and the proposed scheme with optimized task allocation provides a notable gain as compared to all the benchmark schemes.

In Fig. 10, we plot the average task ratio c_k assigned to ENs versus the fronthaul capacity C_F for $N_U \in \{2, 4\}$, $N_E = 2$, $n_{E,i} = 1$, $W = 100$ MHz and $F_{E,i} \in \{0.1, 0.5\} \times 10^{10}$. The task ratio variables are obtained from the proposed algorithm in Sec. IV-D. We observe from the figure that, as the fronthaul capacity C_F increases, more tasks are assigned to CP due to reduced fronthaul latency. Similarly, as the ENs are equipped with stronger computing power $F_{E,i}$, they process a larger portion of tasks. Moreover, increasing the number N_U of UEs results in smaller ratios c_k, since the ENs with limited computing power offload more tasks to the CP when N_U is larger.

VI. Conclusions

We have studied the design of collaborative cloud and edge mobile computing within a C-RAN architecture for minimal end-to-end latency. We have tackled the joint design of computational resource allocation and C-RAN signal processing strategies with the goal of minimizing end-to-end latency required for completing the computational tasks of all the participating UEs in the network. To tackle the non-convex

optimization problem, we have applied FP and matrix FP. Via extensive numerical results, we have validated the convergence of the proposed optimization algorithms, the performance gain of C-RAN architecture as compared to D-RAN, and the impact of optimized computational resource allocation of collaborative cloud and edge computing. As future work, we mention the extension to collaborative AR [13], heterogeneous C-RAN and mobile computing integrated systems [42]–[44], the robust design with imperfect CSI [45], and the energy-efficient design [3], [4] for energy-limited mobile UEs. Also, it would be relevant to verify the effectiveness of the proposed algorithms by deriving a tight lower bound on the optimal latency values.

REFERENCES

[1] H. T. Dinh, C. Lee, D. Niyato and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[2] M. Satyanarayanan, P. Bahl, R. Caceres and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Comput.*, vol. 8, no. 4, pp. 14–23, Oct. 2009.

[3] S. Sardellitti, G. Scutari and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[4] A. Al-Shuwaili, O. Simeone, A. Bagheri and G. Scutari, "Joint uplink/downlink optimization for backhaul-limited mobile cloud computing with user scheduling," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 4, pp. 787–802, Dec. 2017.

[5] T. X. Tran, A. Hajisami, P. Pandey and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 54–61, Apr. 2017.

[6] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd.quart., 2017.

[7] S. Xiao, C. Liu, K. Li and K. Li, "System delay optimization for mobile edge computing," *Future Generation Computer Systems*, vol. 109, pp. 17–28, Aug. 2020.

[8] J. Ren, G. Yu, Y. He and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5031–5044, May 2019.

[9] D. Huang, P. Wang and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 1991–1995, Jun. 2012.

[10] K. Kumar and Y. H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computers*, vol. 43, no. 4, pp. 51–56, Apr. 2010.

[11] G. Scutari, F. Facchinei and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–Part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, Apr. 2017.

[12] G. Scutari, F. Facchinei, L. Lampariello, S. Sardellitti and P. Song, "Parallel and distributed methods for constrained nonconvex optimization–Part II: Applications in communications and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1945–1960, Apr. 2017.

[13] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 398–401, Jun. 2017.

[14] S. Jeong, O. Simeone and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, Mar. 2018.

[15] J. Kang, O. Simeone, J. Kang and S. Shamai (Shitz), "Control-data separation with decentralized edge control in fog-assisted uplink communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3686–3696, Jun. 2018.

[16] Y. Wu, K. Ni, C. Zhang, L. P. Qian and D. H. K. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12244–12258, Dec. 2018.

[17] O. Simeone, A. Maeder, M. Peng, O. Sahin and W. Yu, "Cloud radio access network: Virtualizing wireless access for dense heterogeneous systems," *J. Commun. Netw.*, vol. 18, no. 2, pp. 135–149, Apr. 2016.

[18] S.-H. Park, O. Simeone, O. Sahin and S. Shamai (Shitz), "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.

[19] Y. Zhou and W. Yu, "Fronthaul compression and transmit beamforming optimization for multi-antenna uplink C-RAN," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4138–4151, Aug. 2016.

[20] S.-H. Park, O. Simeone and S. Shamai (Shitz), "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[21] A. Garcia-Saavedra, G. Iosifidis, X. Costa-Perez and D. J. Leigh, "Joint optimization of edge computing architectures and radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2433–2443, Nov. 2018.

[22] S. Yang, N. He, F. Li, S. Trajanovski, X. Chen, Y. Wang and X. Fu, "Survivable task allocation in cloud radio access networks with mobile edge computing," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1095–1108, Jan. 2021.

[23] Q. Zing, L. Gui, F. Hou, J. Chen, S. Zhu and F. Tian, "Dynamic task offloading and resource allocation for mobile-edge computing in dense cloud RAN," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3282–3299, Apr. 2020.

[24] L. Yang, J. Cao, H. Cheng and Y. Ji, "Multi-user computation partitioning for latency sensitive mobile cloud applications," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2253–2266, Aug. 2015.

[25] Y. Mao, J. Zhang and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–33605, Dec. 2016.

[26] A. Yousefpour, G. Ishgaki, R. Gour and J. P. Jue, "On reducing IoT service delay via fog offloading," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 998–1010, Apr. 2018.

[27] K. Shen and W. Yu, "Fractional programming for communication systems–Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[28] K. Shen, W. Yu, L. Zhao and D. P. Palomar, "Optimization of MIMO device-to-device networks via matrix fractional programming: A minorization-maximization approach," *IEEE/ACM Trans. Netw.*, vol. 27, no. 5, pp. 2164–2177, Oct. 2019.

[29] T. K. Y. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1458–1461, Oct. 1999.

[30] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

[31] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming," ver 2.0 beta, Sep. 2013. [Online]. Available: http://cvxr.com/cvx.

[32] A. Ben-Tal and A. Nemirovski, Lecture Note of "Lectures on modern convex optimization – 2019," Georgia Institute of Technology, 2019. [Online]. Available: https://www2.isye.gatech.edu/ nemirovs/LMCO_LN.pdf.

[33] H. Inaltekin and S. V. Hanly, "Optimality of binary power control for the single cell uplink," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6484–6498, Oct. 2012.

[34] H. Joudeh and B. Clerckx, "On the optimality of treating inter-cell interference as noise in uplink cellular networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7208–7232, Nov. 2019.

[35] Z. Yang, Z. Ding, P. Fan and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Trans. Commun.*, vol. 58, no. 10, pp. 6484–6498, Oct. 2012.

[36] A. E. Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge University Press, 2011.

[37] G. Rote, "Division-free algorithms for the determinant and the Pfaffian: Algebraic and combinatorial approaches," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2001, vol. 2122, pp. 119–135.

[38] S. Jeon, B. C. Jung, H. Lee and J. Park, "Interference coordination for heterogeneous users in asynchronous fog radio access networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1064–1068, Aug. 2019.

[39] J. Kim and S.-H. Park, "Broadcast coding and successive refinement for layered UE cooperation in multi-user downlink," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 893–896, Jun. 2020.

[40] M. Chen and Y. Hao, "Task offloading for mobile edge computing in software defined ultra-dense network," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 587–597, Mar. 2018.

[41] R. H. Tutuncu, K. T. Toh and M. J. Todd, "Solving semidefinite-quadratic-linaer programs using SDPT3," *Math. Program.*, Series B, vol. 95, pp. 189–217, Feb. 2003.

[42] M. Peng, Y. Li, J. Jiang, J. Li and C. Wang, "Heterogeneous cloud radio access networks: a new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.

[43] Q. Pham and et al., "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020.

[44] Y. Ma, H. Wang, J. Xiong, J. Diao and D. Ma, "Joint allocation on communication and computing resources for fog radio access networks," *IEEE Access*, vol. 8, pp. 108310–108323, 2020.

[45] E. A. Gharavol and E. G. Larsson, "The sign-definiteness lemma and its applications to robust transceiver optimization for multiuser MIMO systems," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 238–252, Jan. 2013.